The Frugal Inference of Causal Relations

Malcolm Forster, Garvesh Raskutti, Reuben Stern, Naftali Weinberger

Abstract

Recent approaches to causal modeling rely upon the Causal Markov Condition, which specifies which probability distributions are compatible with a Directed Acyclic Graph (DAG). Further principles are required in order to choose among the large number of DAGs compatible with a given probability distribution. Here we present a principle that we call *frugality*. This principle tells one to choose the DAG with the fewest causal arrows. We argue that frugality has several desirable properties compared to the other principles that have been suggested, including the well-known Causal Faithfulness Condition.

1 Introduction

1.1 The Causal Markov Condition

2 Faithfulness

3 Frugality

3.1 Basic Independencies and Frugality

3.2 General Properties of DAGs Satisfying Frugality

3.3 Connection to minimality assumptions

4 Discussion: Frugality as a Parsimony Principle

5 Conclusion

1 Introduction

It is almost a cliché to say that causation is not the same as correlation. Nevertheless, there is a widely held principle, known as the principle of the common cause (PCC), that does state a connection between causation and correlation. It says that there is no correlation without some causation. More specifically, it says that if two events are correlated (in the sense of being probabilistically dependent on each other), then either one event causes the other or both events are effects of some common cause. This principle helps address the problem of how to use an empirically grounded probability distribution over a space of possible events to infer something about causation.

But there are two well-known limitations of the PCC. (1) It has a rather narrow range of applications. At best, it tells us what to infer given a correlation between two events; it does not tell us anything about what to infer when the probability distribution ranges over more events. (2) Even when the principle is applicable, it does not allow us to infer anything very specific about causation. As stated above, it leaves open three possible causal hypotheses, and does not tell us which of them should be preferred. Even worse, the correlation between two events may be produced by a combination of two of these three hypotheses. It is possible that one event causes another and that these two events share a common cause. In this case, there is a direct causal relationship between every pair of events. In the causality literature, this kind of causal hypothesis is often referred to as *complete* Popper (1959).

Recent work on causality is based on what has become known as the Causal Markov Condition (CMC) (see e.g. Glymour *et al.* (1987), Kiiveri *et al.* (1984), Pearl (2000), Spirtes and Zhang (2014)). This principle entails (one formulation of) the PCC as a special case. But the CMC is far more general than the PCC, and is applicable to causal hypotheses involving any number of events. Nevertheless, the CMC has the same limiting property as the PCC: it always fails to rule out complete causal hypotheses. This means, as Zhang (2013) points out, that the CMC must be supplemented by some version of Occam's razor in order to use

probabilistic information to infer specific causal hypotheses. Zhang discusses three such razors: the causal Faithfulness condition, and two versions of the causal minimality condition. In this paper, we introduce a fourth razor, which we refer to as *frugality*, and argue that it is superior to each of the other razors in some important respects. Frugality states that among the causal hypotheses that are compatible with a probability distribution according to the CMC, one should choose from among the set of hypotheses that posit the fewest causal arrows.

The CMC and the various razors have been used to develop and justify search algorithms for choosing among a set of causal hypotheses based on knowledge of a probability distribution. The following paper complements that of Raskutti and Uhler (2006), who present the virtues of search algorithms that seek the most frugal model. Here, however, we do not aim to defend any particular search algorithm. The choice of a search algorithm depends on factors other than (and in addition to) the plausibility of the algorithms' underlying assumptionsâĂŤ for example, on computational and statistical tractability. Here our focus is on the assumptions themselves.

We argue that frugality is preferable to the other razors insofar as it provides a principled basis for choosing among causal hypotheses compatible with the CMC. Frugality specifies a criterion for ranking models in terms of their simplicity âĂŞ- i.e. number of causal arrows âĂŞ and specifies that one should choose from among the simplest models. As we will see, the other proposed razors do not work like this because they do not yield a ranking over the set of models that are compatible with the CMC. For example, the Causal Faithfulness Condition does not specify a basis for ranking models, but rather presents a criterion that models can either succeed or fail to meet. The minimality conditions allow for orderings over subsets of models, but do not allow for a global ordering over the full set of models. The use of these three razors will often lead one to choose simpler models, but only frugality dictates that one should choose these models *because* they are simpler.

Of the three razors, the Causal Faithfulness Condition (CFC) has by far received the most attention in the philosophical literature. A causal model is faithful to a probability distribution if all of the probabilistic independencies in the distribution are entailed by the CMC. The CFC states that the true causal model is faithful to the true probability distribution. Recently Zhang

and Spirtes (2008) have fruitfully explored ways that one can use principles that are logically weaker than faithfulness to choose among causal hypotheses. In cases where faithfulness fails, but these weaker principles are satisfied, it is sometimes possible to use the probability distribution to verify that faithfulness has failed and to avoid choosing the wrong model. Frugality resembles Spirtes and Zhang's principles in being logically weaker than faithfulness. Yet frugality not only enables one to avoid getting the wrong model in cases where faithfulness fails, but also enables the identification of the correct model in many such cases. Moreover, in any case where faithfulness is satisfied, the true model satisfies frugality as well. Frugality thus never fails where faithfulness succeeds, but sometimes succeeds where faithfulness fails.

The remainder of this paper is organized as follows. In section 1.1, we introduce the Causal Markov Condition. Readers who are already familiar with the CMC should feel free to skip to Section 2. In section 2 we give an overview of the literature on the Causal Faithfulness Condition (CFC). Section 3 introduces frugality and compares it to Zhang's three versions of Occam's razor. In section 4, we present frugality's virtues as a model selection principle. Section 5 concludes.

1.1 The Causal Markov Condition

In the following, we will follow standard practice in representing causal hypotheses using directed acyclic graphs, defined as follows:

Definition 1 *A directed acyclic graph (DAG) specifies a fixed set of variables, plus a set of arrows between pairs of these variables, such that there are no cycles when you follow the arrows from tail to tip in an unbroken sequence (a directed chain, or directed path). When the arrows in a DAG are interpreted causally, 'acyclic' implies that no variable is directly or indirectly a cause of itself.*

Note that DAGs do not talk about causal relationships between events. Rather, they posit causal relationships among variables, where a variable can be thought of as a set of mutually exclusive events. When two events are correlated, each event and its negation is viewed as a dichotomous variable.

To interpret a DAG causally is to treat its arrows as representing direct causal relationships between variables such that the value of each variable is a function of the values of its direct causes in the DAG. To illustrate, consider a case in which whether a bell rings causally depends on the outcomes of the flips of two fair coins. [1] The bell rings if and only if both coins are heads or both are tails. We can represent this case in a DAG using three dichotomous variables: one for each coin flip (values: heads, tails) and one for whether the bell rings (values: yes, no). Knowledge of the values of any two of the variables is sufficient for inferring the value of the third. Accordingly, any two variables are correlated conditional on the value of the third (although no pair of variables is unconditionally correlated). To say that one can know the value of any variable given those of the others is not, of course, to say that each variable causally depends on the other two. As we have presented the case, the ringing of the bell depends on the outcomes of the coin tosses, and there are no other causal dependencies amongst these variables. The correct DAG is therefore that given in figure 1. In discussing DAGs, we will refer to a variable's direct causes as its *parents*, and to any causally downstream variables as its *descendants*. Accordingly, the coin flip variables are parents of B and B is a descendant of each. By convention, every variable is considered to be a descendant of itself.

The relationship between the causal hypothesis represented by a DAG and the probability distribution over the variables in the model is given by the Causal Markov Condition:

Definition 2 (Causal Markov condition (Spirtes *et al.* (2000))) *Given a DAG G over variable set* $\mathbf{V}$ *and probability distribution* $\mathbb{P}$ *over* $\mathbf{V}$*, G and* $\mathbb{P}$ *satisfy the Causal Markov Condition if and only if any variable X in* $\mathbf{V}$ *is probabilistically independent of its non-descendants given its parents.*

[1] Cases like this in which three events are pairwise uncorrelated, but where each event is correlated with the combination of the others are instances of *Bernstein's paradox*. Some philosophers Uffink (1999) have considered such cases to be troublesome for the principle of the common cause, and a test case for generalizations of it. The following discussion reveals that such cases are neither counterexamples to the PCC nor to the stronger CMC. Rather, they are failures of faithfulness (see section 2).

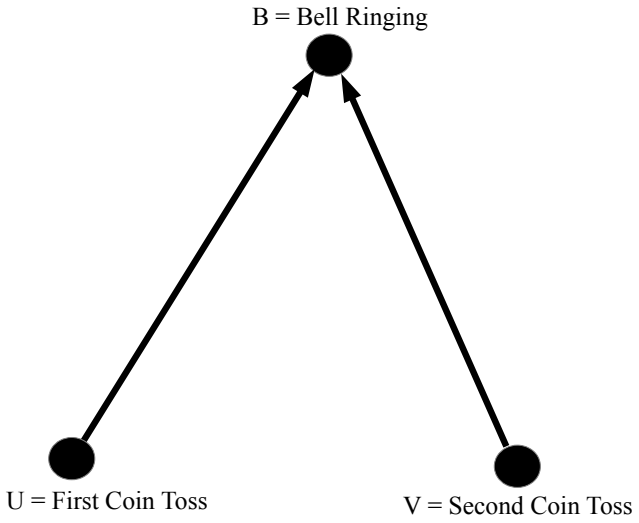

Figure 1: Causal diagram (DAG) for bell example.

In our example, the coin flip variables have no parents, so according to the CMC they are uncorrelated conditional on the variable set that does not include B. In order for the CMC to be plausible, one must assume that the DAG does not omit any common causes of the variables in the model. If there were an omitted common cause of the variables U and V in figure 1, we could not justifiably assume that they were uncorrelated. In what follows, we will assume that the models we present include all common causes of variables in the model and that the true DAG over these variables satisfies the CMC. We further assume that we have knowledge of the true probability distribution and in particular, that we know which variables are conditionally dependent and independent.

As mentioned earlier, the CMC entails a version of the PCC – namely, that if variables X and Y are correlated, then either X (directly or indirectly) causes Y, Y (directly or indirectly) causes X, or X and Y are joint effects of some common cause. The CMC [2] is more general

[2]Though we follow others (see e.g. Spirtes and Zhang (2014), Zhang (2013)) in limiting our discussion to “causally sufficient" variable sets – i.e. variable sets for which it is the case that every common cause of any two or more variables in V is in V – we believe it would be worthwhile to consider applications of frugality in contexts where causal sufficiency is not assumed.

than the PCC. Notably, while the PCC applies only to the correlation between two variables, CMC is applicable to variable sets of any size. On either principle, a complete graph in which every pair of variables is connected by a causal arrow is compatible with any probability distribution. The PCC requires that if X and Y are correlated, then they are causally related, and thus it cannot be falsified by a DAG in which all variables are causally related. According to the CMC, a complete graph entails no conditional independencies, since for every pair of variables in a complete graph, one is a direct descendent of the other. Therefore, the CMC, too, can never rule out the complete DAG. This underscores an important point. Though the CMC tells us that some arrows *must* be included in a given DAG, it never tells us which arrows *should not* be included in a given DAG.

Below we make precise an important conceptual relationship between the number of missing edges in a graph (as compared to the complete graph) and the set of conditional independencies entailed by that graph according to the CMC. Roughly, when one removes an edge from a DAG, the resulting DAG entails more conditional independencies. DAGs that entail more conditional independencies rule out more probability distributions, and are thus more informative. Complete DAGs are minimally informative, since they are compatible with all probability distributions. The fact that the CMC never eliminates the complete graph reveals that it does not require models to be informative. The need for principles other than the CMC can thus be motivated on the grounds that scientists do not aim merely to produce theories that are true and general - logical truths are both - but to produce theories that are informative. Below we defend frugality on the grounds that it requires one to choose a maximally informative model.

In what follows, we refer to the graphs compatible with the probability distribution according to the CMC as Markovian DAGs. For DAGs with more than a few variables, the set of Markovian DAGs is large (see e.g. Raskutti and Uhler (2006)), so the CMC by itself does not enable one to say much about which DAG is the true one. Further principles are therefore required to choose among the Markovian DAGs for a given distribution. In the following section, we review the best known such principle âĂŞ- the Causal Faithfulness Condition.

2 Faithfulness

Of the principles for choosing among the set of Markovian DAGs, the Causal Faithfulness Condition (CFC) (see e.g. [3, 11, 16]) is the most discussed in the philosophical literature. A DAG is faithful to a distribution if it entails all of the conditional independencies (CIs) found in that distribution. More precisely:

Definition 3 (Faithfulness) *Given a set of variables V, a DAG G, and a probability distribution* $\mathbb{P}$ *defined on these variables, G is* faithful *to* $\mathbb{P}$ *if and only if there are no conditional independencies in* $\mathbb{P}$ *that are not entailed by G using the CMC.*

The *Causal Faithfulness Condition* (CFC) states that the true DAG is faithful to the true distribution. Of course, we do not always know which DAG is the true one. When we are trying to determine which DAG is the true one, assuming the CFC amounts to rejecting any candidate DAG that is not faithful to the distribution.

While the CMC requires that a DAG *not* entail any CI that is not in the probability distribution, the CFC requires that *all* of the CI relations found in the distribution are entailed by G. The CFC significantly shrinks the set of candidate Markovian DAGs.

It is well known that the faithfulness assumption is violated in cases where there are two causal paths with equal and opposite effects. To give Hesslow's canonical example Hesslow (1976), taking birth control pills influences one's risk of thrombosis via two paths (figure 2). The pills decrease one's chances of becoming pregnant and pregnancy itself is a cause of thrombosis. Additionally, the pills increase one's risk of thrombosis via a path unrelated to pregnancy. If the positive and negative effects cancel out exactly, then taking birth control pills will be uncorrelated with thrombosis. Since the DAG does not entail that any of the variables will be uncorrelated, this example is a failure of faithfulness.

Before proceeding to evaluate the CFC, it is necessary to say more about causal inference from CIs. In general, CIs do not allow one to discover the unique correct DAG G^*. If two DAGs entail all and only the same CI statements according to the CMC, we say they are in the same *Markov equivalence class*. It is not possible to use CIs to distinguish between DAGs that are in the same Markov equivalence class. The goal of causal inference from CIs is therefore

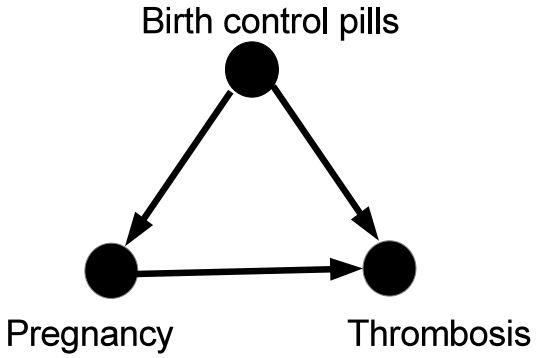

Figure 2: Hesslow's original example: Path from pills to pregnancy to thrombosis cancels out exactly the edge from pills to thrombosis.

to discover the Markov equivalence class of the true DAG, denoted by $\mathcal{M}(G^*)$.

Since the CMC alone never rules out a complete DAG, it is insufficient for discovering $\mathcal{M}(G^*)$ in any case where G^* is not complete. In contrast, CMC and CFC combined are sufficient for discovering $\mathcal{M}(G^*)$.

Theorem 1 (Spirtes *et al.* (2000)) *Assuming both CMC and CFC, the Markov equivalence class $\mathcal{M}(G^*)$ can be identified using the conditional independence statements of probability distribution $\mathbb{P}$.*

Recently, one of the primary roles of the CFC has been in proofs justifying the asymptotic correctness of search algorithms such as PC and FCI. [3] Note that although CMC and CFC is

[3]Searching algorithms can broadly be divided into two classes of methods (or a hybrid of both): constraint-based methods and score-based methods. Constraint-based methods include the traditional SGS and PC algorithms and the more recent FCI algorithm. The CFC is the weakest known sufficient condition for constraint-based methods since it ensures that all conditional independence statements or 'constraints' are respected by the data. Score-based methods include greedy equivalence search, hill-climbing and the more recent penalized ℓ_0-based maximum likelihood method. They operate on the principle that each DAG is assigned a score and the highest scoring graph is selected. There is considerably less work on

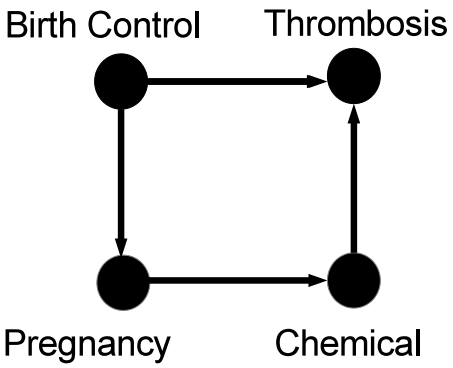

Figure 3: Hesslow's modified example: Path from pills to pregnancy to chemical to thrombosis cancel out exactly with edge from pills to thrombosis.

sufficient for proving the correctness of these algorithms, the algorithms do not *presuppose* the CFC. That is, these algorithms may output a model that does not satisfy the CFC. Accordingly, arguments showing that the CFC is false or unjustified do not directly undermine the reliability of these algorithms, which may in some cases output a true model that violates faithfulness. Here we do not address the question of which algorithms should be used in searching for the correct causal model. Our aim to is to evaluate whether faithfulness (and other so-called razors) provides a principled basis for choosing among models satisfying the CMC.

The thrombosis example is paradigmatic of the failures of faithfulness that have received the most attention in the philosophical literature. It involves three nodes arranged into a

necessary and sufficient conditions for score-based methods, however van de Geer and Bühlmann introduce the permutation beta-min condition for their ℓ_0-based method. The virtues of frugality vs faithfulness in the context of searching algorithms is discussed in Raskutti and Uhler (2006), where it is shown that (i) frugality out-performs faithfulness in terms of learning causal hypotheses if we ignore computational considerations; and (ii) frugality is a sufficient condition for a score-based method known as the SP algorithm introduced in Raskutti and Uhler (2006) and is asymptotically equivalent to the permutation beta-min condition introduced by van de Geer and Bühlmann. For a more thorough discussion see Raskutti and Uhler (2006).

triangle. There are other failures of faithfulness that do not involve triangles. Suppose we modify Hesslow's thrombosis example to include an additional measured variable on the path between pregnancy and thrombosis. This variable could plausibly be a chemical that increases in the bloodstream when a woman gets pregnant (figure 3). If the two paths cancel, this case would also be a failure of faithfulness, but the case is interestingly different than the variant involving the triangle. In the four-node example, there is no DAG that is faithful to the probability distribution. Such failures are called *detectable* failures of faithfulness.

Definition 4 *If a set of variables V and a probability distribution* $\mathbb{P}$ *defined on these variables is such that there is no DAG that is faithful to* $\mathbb{P}$*, then we say that* $\mathbb{P}$ *exhibits a* detectable *failure of faithfulness.*

Such failures are detectable in the sense that it is in principle possible to determine that no DAG is faithful to the distribution and thus to avoid choosing a false faithful DAG. In contrast, the triangle case is an *undetectable* failure of faithfulness, since there is a DAG that is faithful to the probability distribution (fig. 4). When there is an undetectable failure of faithfulness, assuming CFC will lead one to choose the false faithful DAG and nothing in the probability distribution will indicate that a failure of faithfulness has occurred.

All undetectable failures of faithfulness involve triangles. Zhang and Spirtes (2008) make this precise using the notion of triangle faithfulness. Their definition makes use of several notions we must define. X_1 and X_2 are *ad jacent* ($X_1 - X_2$) just in case there is an edge going in either direction between X_1 to X_2. A *path* from X_1 to X_n is an ordered set of variables such that one can get from X_1 to X_n via a set of adjacencies and in which no variable appears twice. Given a path $X_1 - X_2 - X_3$, X_2 is a collider just in case $X_1 \rightarrow X_2 \leftarrow X_3$ and a non-collider otherwise. Here is the definition of triangle faithfulness.

Definition 5 (Triangle faithfulness condition) *[Zhang and Spirtes (2008)] Let* X_1, X_2*, and* X_3 *be any three variables that are all adjacent in* G^**.*

1. *If* X_2 *is a non-collider on the path* $X_1 - X_2 - X_3$ *then in the true distribution* X_1 *and* X_3 *are dependent conditional on any set of other variables that* does not contain X_2.

2. *If X_2 is a collider on the path $X_1 - X_2 - X_3$ then in the true distribution X_1 and X_3 are dependent conditional on any set of variables* that contains X_2.

A *failure* of triangle faithfulness is a case in which X_1, X_2 and X_3 are adjacent and either condition 1 or 2 fails. Triangle faithfulness is weaker than CFC. In the 3-node Hesslow case, triangle faithfulness fails, since birth control and thrombosis are uncorrelated conditional on pregnancy, which is a non-collider on the path from birth control to pregnancy to thrombosis. The following theorem is provable:

Theorem 2 (Zhang and Spirtes (2008)) *Assuming that there are no failures of triangle faithfulness, either G^* satisfies the CFC or no DAG satisfies the CFC.*

In the 4-node example, there is no triangle and therefore no failure of triangle faithfulness. When the probability distribution is the one produced by the canceling paths in figure 3, theorem 2 entails that there is no DAG that is faithful to this probability distribution. Additionally, the bell example we used to introduce the CMC also exhibits a detectable failure of faithfulness (see e.g. Zhang and Spirtes (2008)).

The distinction between detectable and undetectable failures of faithfulness reveals that the faithfulness condition has testable implications. Since the faithfulness condition states that the true DAG is faithful to $\mathbb{P}$, it entails that there is some DAG that is faithful to P. With detectable failures of faithfulness, one can in principle show that there is no DAG that is faithful to the distribution and that faithfulness fails. Zhang and Spirtes (2008) and Spirtes and Zhang (2014) have presented two algorithms for finding detectable failures of faithfulness. These algorithms are âĂIJconservativeâĂİ in the sense that when they find such failures they mark some of the edges in the graph as "unknown".

Zhang and Spirtes show how by testing for detectable failures of faithfulness, one can sometimes avoid accepting the wrong model. A greater victory would be to find the correct model (or, to be precise, the Markov equivalence class of the correct model). In the following section we propose an alternative to the faithfulness condition. This condition identifies the correct model even in some cases where no model is faithful to the distribution.

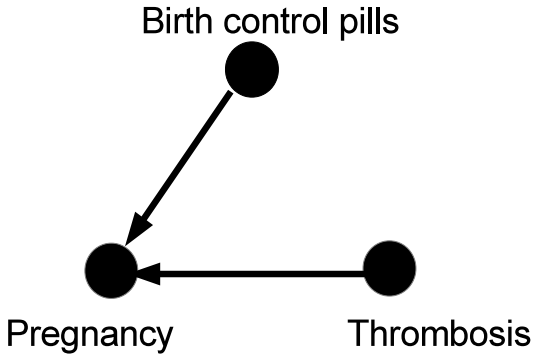

Figure 4: Original Hesslow's example: Faithful DAG that is not the true DAG.

3 Frugality

Importantly, for the 3-node example that violates the CFC in an undetectable way (fig. 2), there exists a DAG with fewer edges (2) that does satisfy the CFC (see fig. 4). Zhang and Spirtes (2008) note that DAGs satisfying the CFC tend to have a small number of edges compared to other Markovian DAGs. We use this idea to develop a different principle for inferring the Markov equivalence class of the true DAG.

A DAG G is *more frugal* than DAG G' if G has fewer edges than G'. Maximally frugal DAGs use only as many edges as are necessary in order to satisfy the CMC. Complete DAGs always trivially satisfy CMC in virtue of not implying any CI relations. But responsible DAG-builders soon learn that the cost of adding edges is to reduce the testable implications of the model and they condemn such exorbitance. Following their lead, we present the *Frugality Condition:*

Definition 6 (*Frugality Condition:*) *Given a probability distribution $\mathbb{P}$ on V, the true DAG over V is in the set of maximally frugal DAGs for V.*

Frugality fails when the true DAG is not in the set of maximally frugal DAGs. Here we present this as a condition on the true DAG, but one could avoid reference to the true DAG by adopting the following methodological substitute (cf. Zhang (2013)):

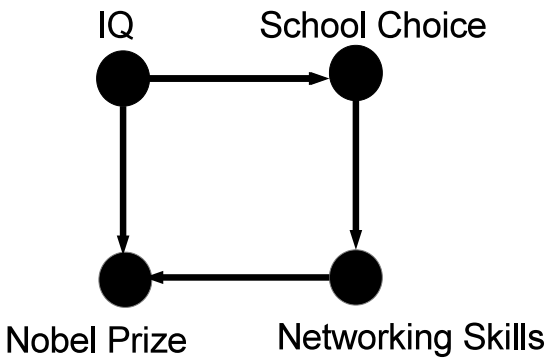

Figure 5: Nobel Prize example: IQ and School choice independent due to path cancellation.

Definition 7 (*Methodological Frugality:*) *Given a probability distribution $\mathbb{P}$ on V, reject all causal structures that are not in the set of maximally frugal DAGs for V.*

We will say that a DAG *satisfies* frugality when it is in the set of maximally frugal graphs and that it *uniquely satisfies* frugality when its Markov equivalence class is the only one in that set. If one does not assume the CMC, then the graph with no edges always uniquely satisfies frugality. Of course, here we are only considering Markovian graphs. To say that a Markovian graph satisfies frugality is simply another way of saying that it is in the set of sparsest Markov representations Raskutti and Uhler (2006). We introduce the new term merely for ease of explication.

Frugality is weaker than faithfulness, since the frugal graph need not entail all of the CI relations in $\mathbb{P}$. Here's an example of where the conditions come apart. Presumably, having a high IQ increases one's chances of winning a Nobel Prize. Aside from being intelligent, it also helps if one has good networking skills. Finally, imagine that intelligent people are more likely to go to the types of schools that encourage one to develop networking skills and these schools are somewhat effective in instilling such skills. The DAG for this example is given in fig. 5.

An interesting feature of this graph is that even if there were no causal path between IQ and networking skills, IQ and networking ability would still be correlated conditional on being a Nobel laureate. In fact, they would be negatively correlated. This accounts for our judgment

that if we meet a Nobel laureate and find her to be totally socially inept, this leads us to think that she must be *really* smart. More generally, conditioning on a common effect (i.e. a collider) induces correlation and this is known in the social sciences as endogenous selection bias Elwert and Winship (2015). Accordingly, if IQ did not influence school choice, IQ and school choice would be negatively correlated among Nobel laureates. Of course, IQ does influence school choice and this opens up the possibility for a type of cancellation different from the one in the Hesslow examples. The negative correlation between IQ and school choice among Nobel laureates might exactly counterbalance the positive effect of IQ on school choice due to the direct causal arrow between them. Consequently, IQ and school choice would be independent conditional on winning the prize. Assuming this were the case, the corresponding probability distribution would contain all of the CIs that are entailed by the DAG in fig. 5 plus the additional one that results from the cancellation.

In the presented case, there is no DAG faithful to the probability distribution, but the correct DAG uniquely satisfies frugality. There is no DAG faithful to the distribution, since no causal structure can account for the conditional independence of IQ and school choice. For example, if you remove the arrows going into and out of the node for school choice, then IQ and school choice would be uncorrelated conditional on being a Nobel winner. But this DAG is inadequate, since it entails the false conclusion that school choice is uncorrelated with any of the other variables. Nevertheless, the true DAG satisfies frugality, since there is no DAG with fewer edges that satisfies the CMC. Using frugality, we can replace a detectable failure of the CFC with a successful recovery of the true DAG.[4]

It is computationally more difficult to determine that a DAG satisfies frugality than it is to determine that it satisfies faithfulness. To determine whether a DAG is maximally frugal, one must consider all causal orderings of the variables and ascertain that there is none for which there is a DAG with fewer edges.[5] Raskutti and Uhler (2006) have developed an algorithm for finding the maximally frugal graph, but here our concern is with the relationship between

[4]In this example there is only one DAG in the Markov equivalence class of the true DAG.

[5]A variable Y follows variable X in the *partial ordering* determined by a DAG if there is a causal arrow from X to Y, that is, $X \rightarrow Y$. Raskutti and Uhler show how given a specification of the ordering over a set of variables, one find a DAG that satisfies CMC.

frugality and faithfulness rather than with the properties of algorithms that assume one condition or the other. Nevertheless, it is worth mentioning their theorem that for every ordering of nodes there is a unique graph that satisfies the CMC with fewest edges. To find the DAGs satisfying frugality, one needs to find the most frugal Markovian graph for each ordering and see which ones have the fewest edges. It should be clear that such a search will never output a graph that does not satisfy the CMC. This distinguishes Raskutti and Uhler's algorithm from the conservative PC algorithm presented in Zhang and Spirtes (2008), which, for example, outputs a structure that does not obey the CMC in the 4-node Hesslow case[6].

The Nobel case is a detectable failure of faithfulness in which the true DAG uniquely satisfies frugality. As we explain below, one can show that whenever the true graph is faithful it is also maximally frugal . The two cases that remain to be discussed are those in which both frugality and faithfulness fail and those in which frugality succeeds, but the true DAG does not uniquely satisfy frugality.

Whenever there is an undetectable failure of faithfulness, both faithfulness and frugality fail. For example, in the 3-node Hesslow example, the faithful DAG only has two edges, and this false DAG uniquely satisfies frugality. Below we will also present an example of a *detectable* failure of faithfulness where frugality fails.

The 4-node Hesslow example is an example of a detectable failure of faithfulness in which the true DAG satisfies frugality, but not uniquely. In addition to the true DAG, there is an additional DAG with four edges (see fig. 6).

In this case one cannot tell which of the maximally frugal DAGs is correct without further information. Applying the frugality condition to this case yields the result that one of the two maximally frugal DAGs is correct. Similarly, in the bell example frugality selects the three DAGs in which there are only two edges arranged as a collider. This is a great improvement over the CMC alone, which yields 9 possible DAGs. It is also an improvement over the combination of the CMC and the CFC, which yields zero graphs. Even when frugality does

[6]Although the 4-node Hesslow case is a detectable failure of the CFC, the conservative PC algorithm makes assumptions that are weaker than faithfulness, so it does output a result in some cases where there is no faithful graph.

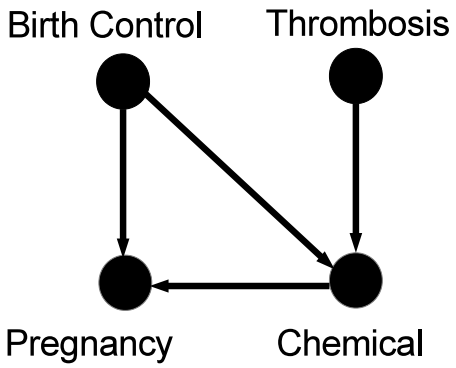

Figure 6: Hesslow's modified example: Frugal DAG that is not the true DAG.

not select a unique DAG, it is still able to significantly reduce the number of DAGs one needs to consider and is therefore very useful.

The three cases we have presented in which faithfulness fails while frugality succeeds are not just failures of CFC, but failures of a weaker condition known as adjacency-faithfulness introduced in Ramsey *et al.* (2006).

Definition 8 *A distribution* $\mathbb{P}$ *satisfies the* adjacency faithfulness assumption *with respect to a DAG* G^* *(with vertices V) if for all pairs of adjacent variables* (X_j, X_k)*,* X_j *is dependent on* X_k *conditioned on any set of variables that excludes* X_j *and* X_k*.*

In the Nobel case, IQ and school choice are independent conditional on being a laureate, even though they are adjacent. In the 4-node Hesslow case, birth control and thrombosis are adjacent, but independent conditional on the empty set. In the bell example, each coin toss is adjacent to the bell's ringing, but is independent of it conditional on the empty set. Thus, even though adjacency-faithfulness is much weaker than CFC, it is nevertheless strong in the sense that it eliminates models that are not eliminated by frugality. Moreover, the conservative PC algorithm Zhang and Spirtes (2008), which was designed in order to avoid certain detectable failures of faithfulness, is only guaranteed to produce a correct result assuming adjacency-faithfulness.

Note that in the 4-node Hesslow case, the two most frugal graphs account for the same

number of CIs. The true DAG accounts for every CI except for the independence of birth control and thrombosis. The other DAG satisfying frugality accounts for every CI except for the independence of birth control and the blood chemical conditional on pregnancy. There is no DAG that is able to account for both of these CIs, which is why the faithfulness condition rules out all models. At least in this case, frugality privileges the models that are able to account for the greatest number of CIs. This suggests a possible connection between a model's having fewer edges and its entailing more CIs. In the following section, we make this connection precise.

3.1 Basic Independencies and Frugality

We have repeatedly emphasized the fact that complete graphs are compatible with any distribution according to the CMC. When one removes edges from a complete DAG, one increases the number of testable implications. Each missing edge may be associated with what we will refer to as a *basic* independency. Given a set of basic independencies corresponding to the missing edges in a graph, one can derive all of the CIs that follow from that graph according to the CMC. In this section, we introduce the concept of a basic independency and explain the connection between such independencies and frugality. For a given set of variables, the number of missing edges will correspond to how frugal a DAG is. It follows that DAGs satisfying frugality entail more basic independencies than those that do not, and are thus more informative.

To introduce the concept of a basic independency, we will once again look at the bell example. In the example, in which neither coin toss causally influences the other, $U \rightarrow V$ is missing and it is this missing arrow that entails the independence of U and V. This is a general property of the CMCâĂŹs application to DAGs: the independencies entailed can be obtained by beginning with a complete causal graph compatible with the causal ordering, then by subtracting one arrow at a time, until we are left only with the arrows in the original DAG. At each step, the CMC entails a single independency, which we shall call a basic independency. In the bell example, when we subtract the arrow $U \rightarrow V$, the probability $P(v|u)$ simplifies to $P(v)$ because the removal of the arrow indicates that the causal mechanism that produces the

value of V does not involve the variable U.

Here is a stepwise procedure for generating a set of basic independencies associated with any graph:

Step 1: Add arrows to complete the DAG. Note that the complete graph entails no independencies, basic or otherwise.

Step 2: Delete one of the added arrows. It doesn't matter which one–the order in which the deletions are made will affect what the basic independencies are, but it will not affect the number, or the total deductive closure of the set of basic independencies obtained.

Step 3: The graph obtained after the deletion of the single arrow will entail a probabilistic independency. Here is how it is calculated. Suppose the deleted arrow is $X \rightarrow Y$. Then the CMC implies that Y is independent of X conditional on the (remaining) parents of Y.

Step 4: Delete another added arrow, and derive another independency (as in Step 3). The independency obtained in Step 3 is still entailed, so now we have two independencies.

Step 5: Keep repeating this until we have deleted all of the added arrows; that is, until we are back to the original DAG.

The independencies obtained by the subtraction of a single arrow are the basic independencies discussed in the proposition above. Collectively, they are basic because all other independencies follow deductively from these using standard graphoid properties Dawid (1979). There is a one-to-one correspondence between the basic independencies and the missing arrows in the graph as illustrated in the proposition below:

Proposition 1 *There is a one-to-one correspondence between basic independencies and the arrows that are missing in any DAG. (Note that given a fixed number of variables, the number of arrows in any complete DAG is fixed, so the number of missing arrows is also fixed.) The exact list of basic independencies depends on both the complete graph with which one begins and the order in which the arrows are subtracted, but their number is always the same. And most importantly, the set of basic independencies (relative to an ordering) will collectively entail all the independencies that are entailed by the DAG according to the CMC.*

To illustrate this proposition, we need to introduce a more complicated example. Consider a set of 4 variables indexed by the numbers $1, 2, 3$, and 4. As a kind of short-hand notation, we

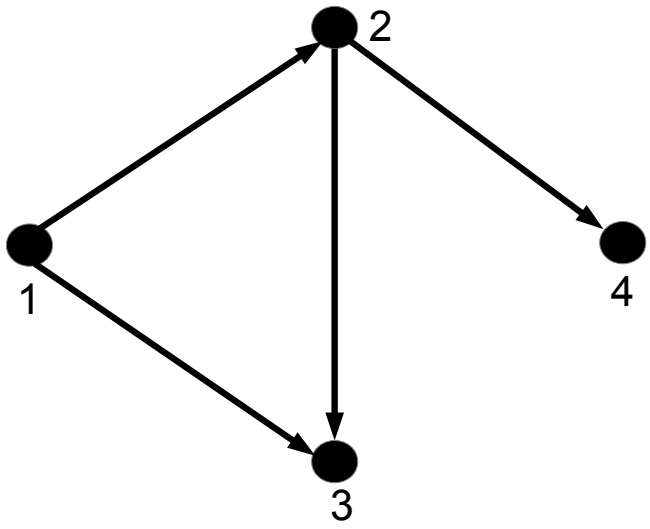

Figure 7: A 4-node directed acyclic graph (DAG).

might write the joint distribution for these variables as $P(1, 2, 3, 4)$. It is a consequence of probability theory alone that their joint distribution decomposes in the following way: $P(1, 2, 3, 4) = P(1)P(2|1)P(3|1, 2)P(4|1, 2, 3)$.

So far no causal assumptions have been made. Now consider the DAG in Fig. 7, and note that the ordering we have chosen, namely $(1, 2, 3, 4)$, is compatible with the causal ordering in the graph. We could just as well have decomposed the probability according to the ordering $(1, 2, 4, 3)$, since this is also compatible with the causal ordering, but we would end up at the same place.

When applied to the causal graph in Fig. 7, the CMC tells us that the probability of each variable in the decomposition can be simplified so that it is conditional on *only* its parents. Any other variables conditioned on in the original decomposition are non-descendants, so they can be omitted. So, for example, $P(4|1, 2, 3)$ simplifies to $P(4|2)$ because 2 is the only parent of 4; 1 and 3 are non-descendants. Thus, the probabilistic independency implied by the CMC in this example is $P(4|1, 2, 3) = P(4|2)$. None of the other factors in the decomposition simplify. So, the final simplified decomposition is $P(1, 2, 3, 4) = P(1)P(2|1)P(3|1, 2)P(4|2)$.

The CMC entails the independence $P(4|1, 2, 3) = P(4|2)$ and all other independencies entailed by the CMC are derivable from this independency. In particular, it entails the two

independencies (a) $P(4|1,2,3) = P(4|2,3)$ and (b) $P(4|2,3) = P(4|2)$. Moreover, it is easy to see that the entailment goes the other way. From (a) and (b), we get $P(4|1,2,3) = P(4|2)$. So the set of independencies {(a), (b)} also has the property that from it, all independencies entailed by the CMC can be derived. However, {(a), (b)} has another important property. Each independency in {(a), (b)} is obtained by the removal of a single arrow. To see this, first add the arrow $3 \rightarrow 4$ to the DAG in Fig. 2. Then we are no longer able to derive (b), but we can still derive (a). So adding a single arrow removes a single independency. Now add $1 \rightarrow 4$ to complete the DAG in Fig. 2. This will prevent us from deriving (a), conforming to the fact that no independencies are entailed by a complete DAG. Again, adding a single arrow removes a single independency. In the reverse direction, begin with the completed graph. Subtract $1 \rightarrow 4$ and obtain (a). Then subtract $3 \rightarrow 4$ to obtain (b). So, removing a single arrow adds a single independency.

As an important consequence of this one-to-one correspondence between missing arrows and basic independences, we have the following proposition.

Proposition 2 *Consider any two DAGs, G_1 and G_2 defined on a fixed set of variables V. G_1 has fewer arrows than G_2 if and only if G_1 entails fewer basic independencies than G_2.*

When comparing DAGs defined over the same set of variables, the number of missing edges will correspond to how frugal a DAG is. Proposition 2 follows immediately: Given any two DAGs defined on the same set of variables, the more frugal DAG entails a greater number of basic independencies. It follows that for a given set of variables, DAGs satisfying frugality entail more basic independencies than those that do not.

We ended the previous section by noting the connection between a model's being more frugal and its entailing more CIs. A natural conjecture is that models satisfying frugality entail more CIs (both basic and otherwise) than those that do not. We do not know whether this is the case (and we would be surprised if it were). What matters, however, is not how many independencies are entailed by a model, but rather how many *independent* independencies it entails. Once one has a set of basic independencies entailed by a model, one already knows everything one needs to know about which probability distributions are compatible with it. This is because all other CIs follow from the set of basic independencies.

Models with more basic independencies have more testable consequences. A given model will entail many propositions, but its testable consequences are those that limit the probability distributions that are compatible with the model. Once one has a complete set of basic independencies, any other CIs do not count as further testable consequences. Since models satisfying frugality entail more basic independencies than those that do not, they are more informative (in the sense of ruling out more probability distributions) than models not satisfying frugality. We will discuss the significance of this point in the discussion section. Before doing so, we will first clarify the properties of DAGs satisfying frugality and the relationship between frugality and other model selection principles for finding the correct DAG.

3.2 General Properties of DAGs Satisfying Frugality

We now present the main result which asserts the properties of DAGs satisfying frugality in relation to faithful DAGs. Parts of the proof rely on the assumption that violations of faithfulness can only occur due to path cancellations in the graph, and not for other reasons. We call this assumption *single-path faithfulness*. A precise definition of single-path faithfulness is given in the appendix. The bell example reveals that frugality is useful even in some cases where the CFC fails for reasons other than cancellation, but we do not yet know the general properties of these cases.

Let $\mathcal{G}_{Fr}(\mathbb{P})$ denote the set of most frugal DAGs satisfying the Markov assumption with respect to $\mathbb{P}$ and $\mathcal{G}_{F}(\mathbb{P})$ denote the set of DAGs that satisfy the CFC with respect to $\mathbb{P}$. We define to be the empty set, or the set that contains no DAGs.

Theorem 3 *(a) Any DAG satisfying the CFC must also lie in the set of most frugal DAGs.*

(b) Whenever the set of DAGs satisfying the CFC is non-empty, it is equivalent to the set of DAGs satisfying frugality

(c) There are examples in which no DAGs satisfy the CFC, whereas the set of DAGs satisfying frugality consists of the true Markov equivalence class.

The combination of (b) and (c) proves that the set of Markovian DAGs G^* where the true Markov equivalence $\mathcal{M}(G^*)$ satisfies the frugality assumption is strictly larger than those satisfying the faithfulness assumption.

3.3 Connection to minimality assumptions

We have so far considered the relationship between frugality and faithfulness, but there are other versions of Occam's razor in the literature. Specifically, there is the so-called minimality principle. As Zhang (2013) notes, there are in fact two versions of the minimality principle, one given by Spirtes, Glymour and Scheines (SGS) and another by Pearl. We will refer to these as SGS-minimality and P-minimality, respectively. In this section we compare frugality to the minimality conditions and prove that it is logically stronger than both.

The SGS-minimality condition asserts that there exists no proper sub-DAG of G^* that satisfies the Markov assumption with respect to P, where a DAG qualifies as a proper sub-DAG of G^* only if (i) it contains fewer directed edges than G^*, and (ii) each of its directed edges is oriented in the same direction as its corresponding edge in G^*. To see that SGS-minimality is weaker than frugality, consider again the true graph in the 3-node Hesslow example and the DAG faithful to the distribution (figure 4 above). As we have seen, the faithful DAG contains fewer edges, so the true graph will not satisfy frugality, but it will be SGS-minimal. The reason for this is that in the faithful graph, the edge goes from thrombosis to pregnancy instead of from pregnancy to thrombosis, so the faithful DAG is not a sub-DAG of the true DAG. A DAG can be SGS-minimal even if there is a graph with fewer edges, provided that either the adjacencies in the smaller graph are different, or that some of them go in the reverse direction. It follows that every graph that is not SGS-minimal will not satisfy frugality, but not that every graph that is SGS-minimal will satisfy frugality. Frugality is stronger, since it rules out a strict superset of DAGs. It bears mention that there will be more SGS-minimal graphs than might appear at first glance. Raskutti and Uhler (2006) show that for every causal ordering for a set of variables, there exists a DAG that satisfies the SGS-minimality assumption.

Pearl's notion of minimality, which we refer to as P-minimality asserts that Markovian DAGs that entail more CI statements are preferred. A DAG fails to be P-minimal just in case

the CIs entailed by the DAG are a proper subset of the CIs entailed by some other Markovian DAG. Stated precisely, let G and G' be two Markovian DAGs and $I(G)$ denote the CI statements entailed by G and $I(G')$ the CI statements entailed by G'. Then G' is *preferred* to G if $I(G) \subset I(G')$ and G' is *strictly preferred* to G if $I(G) \subseteq I(G')$ but $I(G') \not\subset I(G)$. In other words the DAG G' entails a strict super-set of CI statements compared to the DAG G. The P-minimality assumption asserts that no DAG is *strictly preferred* to the true DAG G^*. Zhang (2013) shows that under P-minimality, any violation of faithfulness is detectable since under P-minimality, either G^* satisfies the CFC or no DAG satisfies the CFC. However, as we show in this section, the P-minimality assumption is not as powerful in determining the true Markov equivalence of G^* as is frugality. In particular we show that any DAG satisfying frugality must satisfy P-minimality whereas there are DAGs that satisfy P-minimality, but not frugality. Hence the set of Markovian DAGs satisfying the P-minimality assumption is a superset of those satisfying frugality.

In the 3-node Hesslow example, the true DAG is not P-minimal and the faithful DAG is. The true DAG entails no CIs. The faithful DAG (Figure 4) entails that birth control and thrombosis will be independent conditional on the empty set and since this CI obtains, the model satisfies CMC. This is an example of an undetectable failure of faithfulness in which P-minimality selects the false faithful DAG. Zhang (2013) proves that for all undetectable failures of faithfulness, the faithful DAG will be P-minimal. Accordingly, under P-minimality, any violation of faithfulness is detectable. However, despite the fact that P-minimality and faithfulness favor the same models when failures are undetectable, it is clear that faithfulness is strictly stronger. Given a detectable failure, faithfulness rules out all models, but there is always a set of P-minimal models.

Faithfulness is stronger than either minimality condition. Zhang (2013) further proves that P-minimality is stronger than SGS-minimality. In Theorem 4 we show that the frugal graph will always be P-minimal. The following example reveals that the converse does not obtain. Recall the 4-node DAG involving IQ, school choice, networking skills, and whether someone is a Nobel Laureate where IQ and school choice are independent conditional on meeting a Nobel Laureate as a result of the path cancellation described earlier (fig. 5). Suppose this is

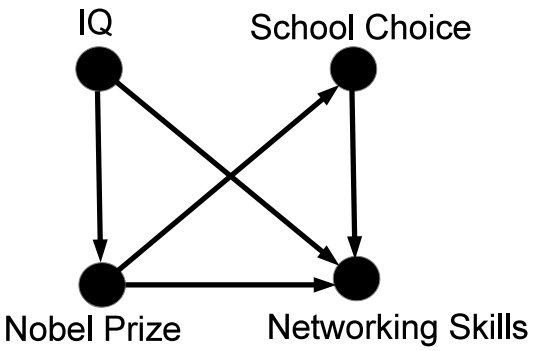

Figure 8: Nobel Prize example: Markovian DAG satisfying P-minimality that is not frugal.

the true DAG. Now consider the DAG with the same nodes containing 5 edges displayed in fig. 8. This DAG entails only the CI statement that IQ is independent of school choice conditioned on Nobel Laureate. Clearly this DAG does not satisfy frugality, however it satisfies the P-minimality assumption since it is Markovian and entails a CI statement that is not entailed by the true DAG or any other Markovian DAG.

The results we have just presented informally were stated more precisely in Raskutti and Uhler (2006). Let $\mathcal{G}_{PM}(\mathbb{P})$ denote the set of Markovian DAGs that satisfy the P-minimality assumption and let $\mathcal{G}_{SGSM}(\mathbb{P})$ denote the set of DAGs that satisfies the SGS-minimality assumption. The following theorem shows that the set of maximally frugal DAGs $\mathcal{G}_{Fr}(\mathbb{P})$ is a strict subset of the set of DAGs satisfying the P-minimality assumption $\mathcal{G}_{PM}(\mathbb{P})$.

Theorem 4 (Theorem 2.8 in Raskutti and Uhler (2006)) *Based on the definitions of $\mathcal{G}_{Fr}(\mathbb{P})$ and $\mathcal{G}_{PM}(\mathbb{P})$:*

(a) $\mathcal{G}_{Fr}(\mathbb{P}) \subset \mathcal{G}_{PM}(\mathbb{P})$.

(b) There exist examples where $G \in \mathcal{G}_{PM}(\mathbb{P})$ while $G \notin \mathcal{G}_{Fr}(\mathbb{P})$.

Combining this result with Theorem 3 and previous results due to Zhang (2013), we have the following nesting property:

Corollary 1 $\mathcal{G}_F(\mathbb{P}) \subset \mathcal{G}_{Fr}(\mathbb{P}) \subset \mathcal{G}_{PM}(\mathbb{P}) \subset \mathcal{G}_{SGSM}(\mathbb{P})$. *If* $\mathcal{G}_F(\mathbb{P}) \neq$ *then* $\mathcal{G}_{Fr}(\mathbb{P}) = \mathcal{G}_F(\mathbb{P})$.

This nesting property proves that the set of DAGs satisfying frugality is a superset of the set of faithful DAGs but a subset of the set of DAGs satisfying the P-minimality assumption.

Before moving on to the discussion, it is worth commenting further on the DAG in figure 8. Suppose that this is the true DAG. If so, then we would have a detectable failure of faithfulness in which frugality fails. Moreover, both the DAG in figure 8 and the frugal DAG would be P-minimal. This shows that using frugality instead of P-minimality opens one up to the risk of eliminating the true DAG from consideration. Despite this possibility, we feel that cases such as the one in figure 8 render P-minimality *less* appealing. P-minimality seems plausible because it appears to privilege DAGs that entail more CI relations found in the distribution. Yet this case reveals that P-minimality will prefer a DAG that entails only a single CI to DAGs that entail many more, provided that no other DAG entails *that* CI. Contrary to to initial appearances, then, P-minimality does not mandate that we should try to increase the total number of entailed CIs.

It is unclear to us what principled defense one could provide for P-minimality.[7] In cases where there is a P-minimal model and there are models that entail a subset of its CIs, the principle entails that we should prefer the P-minimal model, which entails more CIs. However, if there are DAGs that entail CIs that are not in that subset, the principle does not dictate that the P-minimal model just discussed should be preferred to them, even if it entails many more CIs. We understand why one would claim that models that entail more CIs should be preferred. But once one grants that the *number* of CIs is important, why should it further matter whether the CIs entailed by one model are a *superset* of those entailed by the other?

4 Discussion: Frugality as a Parsimony Principle

The set of cases in which the true graph satisfies frugality is a superset of the set of cases in which the true graph satisfies faithfulness. Whenever the true DAG is faithful to the

[7] Pearl (2000) defends P-minimality on the grounds that it prefers simpler theories which are "are more constraining and thus more falsifiable" (46). In the discussion we advocate a similar basis for preferring simpler models. Nevertheless, cases such as the one depicted in figure 8 reveal that it is not in general true that p-minimal models entail more constraints.

distribution, it is also maximally frugal. In at least some cases where no DAG is faithful to the distribution, the true DAG is in the set of maximally frugal DAGs.

As noted, there will be cases in which both faithfulness and frugality are false. In undetectable failures of faithfulness, the two principles deliver the same set of models. If the failure of faithfulness is detectable and frugality fails, then assuming faithfulness will yield no models and assuming frugality will yield false models. Frugality's ability to recover the true DAG in a greater number of cases is thus coupled with an increased risk of getting the wrong DAG. [8] In contrast, by adopting faithfulness one risks missing the true model in cases where it is recoverable (using frugality). Whether one chooses one principle or the other in part depends on whether one places more value on avoiding false positives or false negatives. Spirtes and Zhang's conservative approach implicitly places a higher weight on avoiding false positives. Frugality allows for an alternative approach.

Both frugality and faithfulness play the role of restricting the set of viable models to some subset of the models compatible with the CMC, but this similarity obscures an important difference between the principles. Frugality comes in degrees. For any two models, one can say whether one is more, less, or equally frugal. It follows that there will always be a set of most frugal models. In contrast, a model either is or is not faithful to the probability distribution, which makes it possible for there to be no faithful DAG. [9]

[8]Frugality can be extended to allow for detectable failures if the set of most frugal DAGs contains two DAGs that belong to different Markov equivalence classes. This notion of detectable failures is discussed in Section 2.2 in Raskutti and Uhler (2006) and it is shown that many but not all detectable violations of faithfulness are also detectable violations of frugality.

[9]The claim that faithfulness does not come in degrees may appear at odds with several recent developments in the literature. First Uhler *et al.* (2006), Andersen (2013) have distinguished between exact failures of faithfulness and "near-failures" of faithfulness in which two causal paths nearly cancel. When inferring CIs from finite data, one may not be able to distinguish between near-cancellations and exact cancellations, and the former may be just as problematic as the latter for causal inference. Second, there are results concerning bounds on divergences from faithful surfaces Steel (2013). Neither of these developments are relevant to our claims in the present paper. Here we assume knowledge of the probability

Faithfulness is sometimes referred to as a simplicity principle on the grounds that faithful models have fewer edges than unfaithful ones. We agree that faithful models tend to have fewer edges than unfaithful models and, indeed, we agree that faithful models tend to be simpler. But we deny that faithful models are simpler in virtue of being faithful. Faithfulness is a simplifying assumption, rather than an assumption about simplicity. To the degree that models with fewer edges are simpler, their simplicity derives from their having fewer edges. Equivalently, their simplicity derives from their being more frugal. So to the extent that philosophers have defended faithfulness based on considerations of simplicity, their defenses implicitly rely on the idea that a DAG's simplicity corresponds to its degree of frugality.

Philosophers care about simplicity principles, in part, because they allow us to judge one model as simpler than another. Faithfulness is not like this. Models either are or are not faithful; there is no such relation as âĂŸmore faithful thanâĂŹ."More frugal than" is a relation. Models with fewer edges are more frugal. While it does not make sense to describe one model as being more faithful than another, there are contexts in which one may be inclined to describe one model as more minimal than another (either in the sense specified by Pearl or SGS). Consider SGS-minimality, according to which a DAG is minimal just in case it contains no subgraph that is compatible with CMC. While SGS-minimality (like faithfulness) is an all-or-nothing principle, there is a natural way to rank graphs in terms of closeness to the minimal graph. One might, for example, rank a graph that contains the minimal graph as less minimal than the minimal graph, and a graph that contains that graph as even less minimal. Moreover, since SGS-minimality and Pearl-minimality are strictly weaker than frugality (just as frugality is strictly weaker than faithfulness), it may be tempting to think that SGS-minimality and Pearl-minimality can supply what faithfulness cannot, i.e., that it may be possible to develop a metric of simplicity in terms of minimality that allows for gradations of simplicity.

---

distribution, and thus abstract away from concerns relating to inferring CIs from finite data. Moreover, in considering the CIs that are entailed by a model, we do not rely on any stipulations about the parameters in the model. A model either is or is not faithful, notwithstanding further questions about how we *know* whether it is.

Even if one developed a way of ranking graphs by degree of minimality, the ranking metric would importantly differ from that supplied by considerations of frugality because of its reference to subset relations. SGS-minimality is defined in terms of subsets of directed edges. P-minimality is defined in terms of subsets of conditional independence relations. Criteria defined in terms of subsets of directed edges or CI relations will neither guarantee the minimization of edges nor the maximization of CI relations. For instance, recall that when there is a triangle failure of faithfulness, both the triangle and the collider will count as SGS-minimal even though the collider has fewer edges (because the collider's edges are not oriented in the same direction as the corresponding edges in the triangle). Likewise, in the Nobel Prize case, both the true DAG in figure 5 and the 5-sided DAG in figure 8 are P-minimal, despite the fact that the former entails two CI relations and the latter entails just one (because the CI relations entailed by the 5-sided DAG are not a proper subset of the CI relations entailed by the DAG in figure 5). Insofar as one seeks a criterion that either minimizes edges or maximizes CI statements, the minimality criteria test whether one of these is maximized for only a subset of Markovian DAGs. Frugality is not like this. Frugality says that one should choose the DAG that minimizes edges across all Markovian DAGs.

Though we do not know if frugality optimizes the number of total CI relations, we do know that it maximizes the number of basic independencies, and that all other independencies can be derived from the set of basic independencies. The informativeness of a DAG corresponds to the constraints it places on possible probability distributions. Since more frugal DAGs entail more basic independencies, they imply more constraints than their less frugal counterparts. Frugality may therefore be justified by the principle that, all else being equal, one should prefer models with more testable consequences. As Popper famously argued in Popper (1959), more informative models are preferable because they make more falsifiable claims. Though we do not take a stand on whether Popper is correct, the preference for more informative models is commonplace in philosophy of science. If we prefer more informative models, we should prefer more frugal DAGs.

While frugality ranks models in terms of their informativeness, CFC cannot be used to provide any similar ranking. CFC does not say that one should choose the model that best

matches some feature of the probability distribution. Rather, it says that we should only choose models that are consistent with faithfulness. Instead of presenting a metric that orders Markovian models, Faithfulness stipulates a condition in addition to CMC that models must satisfy.

An analogy clarifies this distinction between CFC and frugality. A shadow is produced by an unobserved object, and one wants to infer properties of the object based on the shape of the shadow. Consider the set of objects that are deemed to be compatible with a given shadow. If we find that our criterion of compatibility does not yield a manageable number of possible objects, we would look for ways to choose between the possibilities based on their ability to capture features of the shadow. Faithfulness is akin to further restricting the kinds of objects that we deem to be compatible with the shadow. Frugality makes no such restriction. According to frugality, there is not some additional dichotomous property (beyond satisfaction of the CMC) that renders a DAG kosher. Rather, by frugality's lights every DAG that satisfies CMC is compared and ranked.

Much of the debate over faithfulness has revolved around whether every true causal model is faithful. It is puzzling why, in the context of model selection, philosophers have been so preoccupied with whether true models are always faithful. Model selection criteria identify some metric that one can use as grounds for selecting between possibly true models. Thus, any suitable model selection criterion must identify some feature other than truth that one can use as grounds for preference over models. As we note above, frugality offers such a feature: informativeness. It is not clear what comparable feature CFC provides. Perhaps there is some feature of faithful models in virtue of which they tend to be preferable. But if so, there should exist some principle invoking that feature, rather than faithfulness itself.

Philosophers often treat faithfulness not as a model selection criterion, but as a convention for reading conditional dependencies off DAGs. The convention we have in mind is that given knowledge of the true DAG, if the CMC does not entail that two nodes are uncorrelated, one should (perhaps defeasibly) assume that they are correlated. Our comments here do not relate to the question of whether such a convention is a good one to adopt in a particular context. What one should infer from the true graph is a different question from that of which inductive

principles one should adopt in the process of finding the true graph. Here we have emphasized the latter question.

Throughout this paper, we have assumed that one knows the probability distribution for each case. In making this assumption, we bypass important statistical questions regarding the inference of the probability distribution from one's data. Finite data sets are "noisy" and it is possible to mistakenly infer that uncorrelated variables are correlated or vice versa. A simulation study conducted in Raskutti and Uhler (2006), suggests that the frugality is especially beneficial in the presence of finite noisy data since the faithfulness assumption is very sensitive to errors made in inferring CI statements. It would be useful to capture the effects of noise on both model selection principles and understand whether one is more or less robust to the effects of noise.

Nothing we have said here shows that faithfulness is not a useful assumption. In practice, it is simple to build algorithms that assume faithfulness and computationally expensive to find the most frugal DAG. Since whenever there is a faithful DAG it is also the most frugal DAG, it may sometimes turn out that the extra expense of finding the most frugal DAGs in cases where there is no faithful DAG may not be pragmatically justified. Our criticism of faithfulness is not that it isn't useful, but rather that it does not provide grounds upon which to rank models as being more or less preferable.

5 Conclusion

In this paper, we have presented a novel principle for choosing among DAGs compatible with a probability distribution according to the Causal Markov Condition. The frugality condition has important advantages over the the Causal Faithfulness, SGS-minimality, and P-minimality conditions insofar as it provides a principled basis for choosing among DAGs compatible with the CMC. One advantage is that it allows for a ranking over the complete set of Markovian graphs. Another advantage is that the ranking of graphs is principled; it is based on their informativeness. The CFC does not share these advantages. The fact that faithfulness does not always provide a ranking (principled or otherwise) is evident from the fact that in some cases there is no faithful model. Moreover, the minimality principles do no better. Because they

only compare subsets of Markovian models, the minimality principles fail to provide grounds for preferring either the models with the fewest edges or the models that entail the most conditional independencies. In contrast, frugality provides grounds for preferring the models that have the fewest arrows and that thereby entail the maximal number of basic independencies. So, if a razor is construed as a principle that provides grounds on which to prefer simpler (or more informative) models, then of the four razors on offer, frugality is most deserving of its title.

Appendix A Appendix

Some appendix text goes here.

Funding

I was funded by a very nice funding institution.

Acknowledgements

Several of our theorems in section 3.1 were proven under the assumption of single-path faithfulness. When single-path faithfulness obtains, all failures of faithfulness are due to cancelling paths. To make this idea precise, we need to introduce the concept of d-separation.

Definition 9 (*d-separation*) *Pearl (1988) A path is d-separated by variable set* $\mathbf{Z}$ *just in case:*

(a) The path contains a triple $i \rightarrow m \rightarrow j$ *or* $i \leftarrow m \rightarrow j$ *such that* m *is in* $\mathbf{Z}$*, or*

(b) The path contains a collider $i \rightarrow m \leftarrow j$ *such that* m *is not in* Z *and no descendant of* m *is in* $\mathbf{Z}$*.*

d-separation is property of paths. Two variables are d-separated by $\mathbf{Z}$ *iff* $\mathbf{Z}$ *blocks all paths between those variables.*

The CMC entails that if DAG D is the true graph, all nodes that are d-separated in D are uncorrelated in the probability distribution $\mathbb{P}$.

Single-path faithfulness is defined as follows.

Definition 10 (Single-path-faithfulness assumption) *A distribution $\mathbb{P}$ satisfies the* single-path-faithfulness assumption *with respect to a DAG $G = (V, E)$ if $i \not\perp\!\!\!\perp j \mid S$ for all triples (i, j, S) with $i, j \in V$ and $S \subset V \setminus \{i, j\}$ such that there is a unique path that d-connects i and j given S.*

This assumption is satisfied for example in the Gaussian setting, since a partial correlation is a weighted sum of all paths between two vertices Steel (2013). We conjecture that the single-path faithfulness assumption is satisfied for all linear models.

Appendix - Proof of Theorem 3

Most of the proofs of Theorems 3 follow almost directly from results in Raskutti and Uhler (2006). We require the following additional definition of a weaker form of faithfulness known as orientation faithfulness introduced in Ramsey *et al.* (2006).

Definition 11 *A DAG $G = (V, E)$ if it satisfies the* orientation-faithfulness assumption *with respect to a distribution $\mathbb{P}$ if: for all variables X_1, X_2 and X_3 where $X_1 - X_2 - X_3$ and all subsets $S \subset V \setminus \{1, 2\}$ such that X_1 is d-connected to X_2 given S,*

$$X_1 \not\perp\!\!\!\perp X_2 \mid X_S.$$

For part (a), we prove the contra-positive. Assume $G = (V, E_G) \notin \mathcal{G}_{Fr}(\mathbb{P})$. Then there exists a Markovian DAG $H = (V, E_H)$ such that $(i, j) \notin E_H$ but $e = (i, j) \in E_G$. Since $(i, j) \notin E_H$, $X_i \perp\!\!\!\perp X_j | X_S$ where S are the parents of X_i and X_j since H satisfies the Markov assumption. This implies that G does not satisfy adjancency faithfulness since for a DAG to satisfy adjacency faithfulness there can only be an edge (i, j) if $X_i \not\perp\!\!\!\perp X_j \mid X_S$ for all $S \subset V \setminus \{i, j\}$. Hence $G \notin \mathcal{G}_F(\mathbb{P})$.

For (b), assume $G \in \mathcal{G}_F(\mathbb{P})$. Then G satisfies adjacency faithfulness and orientation faithfulness. Following the proof of Lemma 2.2 in Raskutti and Uhler (2006), $\mathcal{G}_{Fr}(\mathbb{P})$ is contained in the Markov equivalence class of G. Hence $\mathcal{G}_{Fr}(\mathbb{P}) \subset \mathcal{G}_F(\mathbb{P})$ and combined with (a), $\mathcal{G}_F(\mathbb{P}) = \mathcal{G}_{Fr}(\mathbb{P})$. See the DAG example in Figure 5 for (c).

Name

Institution

Department

Address line 1

Address line 2

E-Mail

References

Andersen, H. [2013]: 'When to expect violations of causal faithfulness and why it matters', *Philosophy of Science*, **80**, pp. 672–683.

Dawid, A. P. [1979]: 'Conditional independence in statistical theory (with discussion)', *Journal of the Royal Statistical Society, Series B*, **41**, pp. 1–31.

Elwert, F. and Winship, C. [2015]: 'Endogenous Selection Bias', *Annual Review of Sociology*.

Glymour, C., Scheines, R., Spirtes, P. and Kelly, K. [1987]: *Discovering Causal Strucure*, Academic Press.

Hesslow, G. [1976]: 'Discussion: Two Notes on the Probabilistic Approach to Causality', *Philosophy of Science*, **43**, pp. 290–292.

Kiiveri, H., Speed, T. and Karlin, J. B. [1984]: 'Recursive causal models', *Journal of the Australian Mathematical Society (Series A)*, **36**(1), pp. 30–52.

Pearl, J. [1988]: *Probabilistic reasoning in intelligent systems*, San Mateo: Morgan Kaufman.

Pearl, J. [2000]: *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Popper, K. [1959]: *The logic of scientific discovery*, Routledge.

Ramsey, J., Zhang, J. and Spirtes, P. [2006]: 'Adjancency-Faithfulness and Conservative Causal Inference', in *Uncertainty in Artificial Intelligence (UAI)*, pp. 401–408.

Raskutti, G. and Uhler, C. [2006]: 'Learning directed acyclic graphs based on sparsest permutations', Tech. rep., Department of Statistics, University of Wisconsin-Madison.

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

Spirtes, P. and Zhang, J. [2014]: 'A uniformly consistent estimator of causal effects under the k-triangle faithfulness assumption', *Statistical Science*, **29**(4), pp. 662–678.

Steel, D. [2013]: 'Geometry of faithfulness assumption in causal inference', *Annals of Statistics*, **41**, pp. 436–463.

Uffink, J. [1999]: 'The Principle of the COmmon Cause Faces the Bernstein Paradoz', *Philosophy of Science*, **66**, pp. 512–525.

Uhler, C., Raskutti, G., Buhlmann, P. and Yu, B. [2006]: 'Homogeneity, selection, and the faithfulness condition', *Minds and Machines*, **16**, pp. 303–317.

Zhang, J. [2013]: 'A Comparison of Three Occam's Razors for Markovian Causal Models', *Brit. J. Phil. Sci.*, **64**, pp. 423–448.

Zhang, J. and Spirtes, P. [2008]: 'Detection of unfaithfulness and robust causal inference', *Minds and Machines*, **18**, pp. 239–271.