addition to directed arcs, MAGs can also contain bi-directed arcs $X \leftrightarrow Y$ (indicative of marginalization) and undirected edges $X \rightarrow\leftarrow Y$ (indicative of selection). The causal sufficiency assumption states that there are no hidden common causes of the observed variables in $\mathcal{G}$, which implies that the distribution over the observed variables still conforms to a Bayesian network. In this article we ignore selection bias (= no undirected edges), but do not rely on causal sufficiency.

The equivalence class $[\mathcal{G}]$ of a graph $\mathcal{G}$ is the set of all graphs that are indistinguishable in terms of (Markov) implied independencies. For a DAG or MAG $\mathcal{G}$, the corresponding equivalence class $[\mathcal{G}]$ can be represented as a partial ancestral graph (PAG) $\mathcal{P}$, which keeps the skeleton (adjacencies) and all invariant edge marks, i.e. tails ($\cdot$) and arrowheads ($\rightarrow\leftarrow$) that appear in all members of the equivalence class, and turns the remaining non-invariant edge marks into circles ($\circ$) (Zhang, 2008). The invariant arrowhead at an edge $A \rightarrow\leftarrow B$ in $\mathcal{P}$ signifies that $B$ is not a cause of $A$. An edge $A \rightarrow B$ implies a causal link $A \rightarrow B$ that is also direct.

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.

Causal discovery procedures

A large class of constraint-based causal discovery algorithms is based directly on the faithfulness assumption: if a conditional independence $X \perp \!\!\!\perp Y \mid Z$ can be found for any set of variables $Z$, 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, the FCI algorithm in conjunction with the additional orientation rules in (Zhang, 2008) is the only one that 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 $X \perp \!\!\!\perp Y \mid Z$ should suffice to orient an unshielded triple $(X, Z, Y)$, 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 $S(\mathcal{G}, \mathcal{D})$ that measures how well a Bayesian network with structure $\mathcal{G}$ fits the observed data $\mathcal{D}$, while preferring simpler networks, with fewer free parameters, over more complex ones. If the causal relations between the variables in $\mathcal{D}$ 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 $\mathcal{B} = (\mathcal{G}, \Theta)$, the likelihood of observing a particular data set $\mathcal{D}$ can be computed recursively from the network. Integrating out the parameters $\Theta$ in the conditional probability tables (CPTs) then results in:

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

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

---

1We follow the standard assumption that Markov equivalence implies statistical equivalence (Spirtes, 2010).