Two minimal uDAGs for a node W that destroys X | Y | Z
(b) (c)(a)
QX
W
V
Z
YWX VZY (b)(a)
(a) (b)
X
W Z
Y
V T
U
X
W Z
Y
V T
U

Figure 1: (a) causal DAG with hidden variables, (b) uDAG with unfaithful X ⊥⊥ T | Z for a node W that destroys X | Y | Z
(b) (c)(a)
QX
W
V
Z
Y
V
T
U
X
W Z
Y
V
T
U

Example 2. With Lemma 4 we infer from Figure 1(b) that X ∥p T | V, W. We also find that Y ∥p V, from which, in combination with Y ⊥⊥ V | Z, we (rightly) conclude that (Z ⇒ Y) OR (Z ⇒ V), see (a).

In general, Lemmas 3 and 4 assert different dependencies for different uDAG members of the same equivalence class. If the uDAG G is optimal, then all in/dependence statements from any uDAG member of the corresponding equivalence class [G] are valid. In that case we can do the inference based on the PAG representation P of [G]. This provides additional information, but also simplifies some inference steps. Again, see (Claassen and Heskes, 2012) for details.

For example, identifying an absent causal relation (arrowhead) X ⊨ 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. Let G be an optimal uDAG to a faithful MAG M, then the absence of a causal relation X ⊨ Y can be identified, if there is no potentially directed path from X to Y in the PAG P of [G].

Proof sketch. The optimal uDAG G is obtained by (only) adding edges between variables in the MAG 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 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 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 Supplement. 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.

3.4 Consistent prior over structures

The computation of p(L|Dx) requires a prior distribution p(M) over the set of MAGs over X. A straightforward solution is to use a uniform prior, assigning equal probability to each M ∈ 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.(4), 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(M) through sampling: generate random graphs over all variables in accordance with the specified characteristics, sample one or more random subsets of variables size K, and compute the marginal structure over that subset. Averaging over structures that are MAG-isomorphs (equivalence classes identical under relabeling) improves both consistency and convergence.

Irrespective of the method, it is essential to ensure the prior 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(“X ⊥⊥ Y”) = 6/25, whereas a uniform distribution over two-node DAGs implies p(“X ⊥⊥ Y”) = 1/3. We can weight uDAG representatives according to the size of their equivalence class.

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 1, except that now