Marginal causal consistency in constraint-based causal learning

Anna Roumpelaki **Giorgos Borboudakis** **Sofia Triantafillou** **Ioannis Tsamardinos**
Computer Science Dept.
University of Crete
Voutes University Campus, Heraklion 70013, Crete

Abstract

Maximal Ancestral Graphs (MAGs) are probabilistic graphical models that can model the distribution and causal properties of a set of variables in the presence of latent confounders. They are closed under marginalization. Invariant pairwise features of a class of Markov equivalent MAGs can be learnt from observational data sets using the FCI algorithm and its variations (such as conservative FCI and order independent FCI). We investigate the consistency of causal features (causal ancestry relations) obtained by FCI in different marginals of a single data set. In principle, the causal relationships identified by FCI on a data set $\mathcal{D}$ measuring a set of variables $\mathbf{V}$ should not conflict the output of FCI on marginal data sets including only subsets of $\mathbf{V}$. In practice, however, FCI is prone to error propagation, and running FCI in different marginals results in inconsistent causal predictions. We introduce the term of marginal causal consistency to denote the consistency of causal relationships when learning marginal distributions, and investigate the marginal causal consistency of different FCI variations.Results indicate that marginal causal consistency varies for different algorithms, and is also sensitive to network density and marginal size.

1 INTRODUCTION

Maximal Ancestral Graphs (MAGs) [12] can represent the causal relationships among a set of measured variables, as well as the conditional independence of their joint probability distribution, in the presence of latent confounders.

Under the causal Markov and Faithfulness assumptions [14], every conditional independence that holds in the distribution can be identified in the graph using the criterion of m-separation. MAGs have several attractive properties: They are closed under marginalization, and they are pairwise Markov: Every missing edge in the graph corresponds to a conditional independence in the distribution.

The **independence model** of a joint probability distribution is the set of conditional independencies entailed by the distribution. The set of MAGs that entail the same independence model define a **Markov equivalence class**. All invariant pairwise features of a Markov equivalence class of MAGs can be represented using a **Partial Ancestral Graph (PAG)** [14]. FCI [14, 15] is the first sound and complete algorithm that identifies a PAG given a data set over a set of possibly confounded variables $\mathbf{V}$ and a test of conditional independence.

Since MAGs are closed under marginalization, FCI can be also used in any subset $\mathbf{V} \setminus \mathbf{L}$ of $\mathbf{V}$ to obtain the corresponding marginal PAG. Naturally, the causal features of marginal PAGs should not contradict those of the PAG on the full set of variables: a disagreement between a PAG and a marginal PAG can only be a result of the sensitivity of FCI to statistical errors. We use the term **marginal causal consistency** to describe the degree of agreement of causal relationships among a PAG and its marginals. To the best of our knowledge, this type of consistency of constraint-based causal discovery has not been examined before. We examine the outputs of FCI [14], order-independent FCI [3], and conservative FCI with the majority rule heuristic for collider orientations [11, 3].

In a simulated setting, we found that the algorithms' consistency is sensitive to network density and number of variables that are marginalized out. To examine whether the consistency of causal relationships correlates with the correctness of the induced causal features, we ranked them according to their frequency of appearance in randomly selected marginals, and compared the resulting AUCs with bootstrapping. While marginal consistency measures the sensitivity of an algorithm to a specific choice of measured variables, bootstrapping measures the sensitivity of an algorithm to the specific sample. Results show that marginal consistency can help identify accurate causal fea-

tures. However, bootstrapping outperforms marginal-based ranking in all cases.

The rest of this paper is structured as follows: Section 2 introduces basic MAG notions and notation. Section 3 defines marginal causal consistency in FCI outputs, presents a sound and complete method for identifying all pairwise causal ancestry relationships in a PAG, and compares internal consistency of outputs of different FCI variations. Section 4.1 describes work related to obtaining confidence estimates for causal ancestry relations. Section 4 describes an algorithm for ranking causal relationships for a given data set, and compares the AUC of the proposed approach to confidence estimates obtained with bootstrapping. Conclusions and future directions are discussed in Section 5.

2 PRELIMINARIES

We use V to denote random variables, and $\mathbf{V}$ to denote a set of variables. A graph $\mathcal{G}$, denoted as $\mathcal{G} = (\mathbf{V}, E)$ is defined over a set of variables $\mathbf{V}$ with edges E. A path p is a sequence of adjacent edges, without repetition. A directed path is a path where all edges are directed and have the same direction. We use $X \dashrightarrow Y$ to denote that there exists a directed path from X to Y (X is an ancestor of Y in $\mathcal{G}$).

We use X $\perp\!\!\!\perp$ Y $|\mathbf{Z}$ to denote variables X and Y to be independent given a set $\mathbf{Z}$. In a graph $\mathcal{G}$ a vertex V is a collider on a path u if and only if there are two distinct edges on u containing V as an endpoint and both are into V. Otherwise, V is a non-collider on U. In a graph $\mathcal{G}$, a triplet $X - V - Y$ is unshielded if X and Y are not adjacent in $\mathcal{G}$.

A Bayesian network is represented by a Directed Acyclic Graph (DAG) $\mathcal{G}$ over a set of variables $\mathbf{V}$ and a joint probability distribution P. A directed edge denotes a direct causal relationship. We assume that the following two conditions hold: the Causal Markov Condition (CMC), which states that every variable is independent of its non-descendants given it's parents, and the Faithfulness Condition (FC), which states that all the conditional independencies that hold in P stem from $\mathcal{G}$ and the CMC.

MAGs are mixed graphs, i.e. they can have both directed and bi-directed edges. A directed edge $X \rightarrow Y$ indicates that X causes Y, while a bi-directed edge $X \leftrightarrow Y$ indicates that X and Y are confounded. Two nodes can be connected with only one type of edge, and ancestry has precedence over confounding: if X causes Y and the two are also confounded, only the directed edge $X \rightarrow Y$ is present in the MAG. MAGs can also be used to represent selection bias via undirected edges. For this paper, we assume no selection bias and only consider MAGs with directed and bi-directed edges.

A path is called uncovered if every consecutive triple on the path is unshielded [15]. Also, a path is potentially directed if it can be oriented into a directed path by changing the circles on the path into appropriate tails or arrowheads.

Under the causal Markov and Faithfulness assumptions [14], the conditional independencies that hold in a joint probability distribution can be identified in the corresponding causal graph according to the graphical criterion of m-separation [12]. Constraint-based methods query the data to identify the independence model, and then try to find the causal MAGs that satisfy all and only the observed independencies. A class of Markov equivalent MAGs, that differ in a subset of edge orientations, will in general satisfy the observed constraints. PAGs have the same adjacencies and all invariant orientations shared by all MAGs in a Markov equivalence class. Specifically, an edge end-point is oriented as an arrowhead ('>') or tail ('-') in a PAG if and only if it is invariant in all MAGs represented by it, and is left as a circle ('o') otherwise.

FCI is a sound and complete algorithm for discovering PAGs from observational data, but it is prone to statistical errors. Several extensions have been proposed that aim to tackle the sensitivity of FCI to error propagation: order independent FCI, conservative FCI and majority rule FCI among others. Order independent FCI (denoted iFCI in the rest of this paper), proposed by [3], outputs a PAG that does not depend on the order the variables are given. Conservative FCI [11] checks all unshielded triplets in the following way: for every unshielded triple $X - Y - Z$ check all subsets of X's potential parents and all subsets of Z's potential parents. If Y is not in any subsets, then orient the triplet as a collider; if Y is in all subsets leave the triplet as a non-collider; otherwise tag the triple as unfaithful. Majority rule FCI (denoted mFCI in the rest of this paper) is slightly less strict than conservative FCI. In this extension, an unshielded triple is oriented as a collider if Y is in less than 50% of the subsets and as a non-collider if it is in more than 50% of the subsets. To avoid unfaithful triples in case of ties, we leave the triple as a non-collider. We examine and compare the marginal consistency of FCI, order-independent FCI and majority rule FCI in simulated data.

3 MARGINAL CAUSAL CONSISTENCY IN PAGS

We now define the problem of marginal causal consistency of a constraint-based algorithm. Intuitively, we are interested in how much marginalizing out variables and rerunning FCI (or a variation) preserves *causal* relationships.

To help define the problem, we use $An_{\mathcal{P}}$ to be the set of all ancestral relationships that hold in the Markov equivalence class $[\mathcal{P}]$ entailed by $\mathcal{P}$. Notice that this set may be different than the set of ancestral relationships $TR_{\mathcal{P}}$ that can be identified directly from $\mathcal{P}$ by taking the transitive closure of the directed edges: It may include additional pairs that

are connected by a different causal path in every member of $[\mathcal{P}]$, so that there is no fully oriented path in $\mathcal{P}$. Figure 1 shows an example where an invariant causal relationship does not correspond to a single directed path in the PAG: while $TR_{\mathcal{P}}$ for the PAG of Figure 1a is the empty set, $An_{\mathcal{P}} = \{(X, Y)\}$.

Identifying $An_{\mathcal{P}}$ is not trivial. [1] use a method to implicitly enumerate all MAGs in a Markov equivalence class represented by $\mathcal{P}$. The same techniques can be used to identify $An_{\mathcal{P}}$, by identifying the invariant ancestral relations present in all such MAGs[1].

However, in this work we show that all causal relationships that are invariant in $\mathcal{P}$ but are not in $TR_{\mathcal{P}}$ correspond to a specific pattern, illustrated in Figure 1. The pattern can be easily found in $\mathcal{P}$ to identify all additional relationships that are causal in $\mathcal{P}$ but not present in $TR_{\mathcal{P}}$. Theorem 3.1 proves soundness and completeness of this rule.

Theorem 3.1 *Let $\mathcal{P}$ be the PAG over a set of variables $\mathbf{V}$ representing the Markov equivalence class of MAGs $[\mathcal{P}]$. Then, if $(X, Y) \notin TR_{\mathcal{P}}$, $X \dashrightarrow Y \in An_{\mathcal{P}}$ if and only if $\exists U, V, U \neq V$ such that*

1. *$\langle X, U, \dots Y\rangle$ and $\langle X, V, \dots Y\rangle$ form uncovered potentially directed paths and*
2. *$\langle U, X, V\rangle$ is an unshielded definite non collider in $\mathcal{P}$.*

Proof ($\Leftarrow$) Let U, V be distinct variables such that $\langle X, U, \dots Y\rangle$ and $\langle X, V, \dots Y\rangle$ are uncovered p.d paths, with $\langle U, X, V\rangle$ an unshielded definite non-collider in $\mathcal{P}$. Let $U \star\!\!\rightarrow X$ in some MAG in $[\mathcal{P}]$, where $\star$ is used as a meta-symbol denoting any plausible orientation. Then $X \rightarrow V$ (since U, X, V form a definite non collider) and $V \dashrightarrow Y$ (since every triple on the path is a definite non collider). Else if $U \leftarrow X$, then $U \dashrightarrow Y$ (since every triple on the path is a definite non collider). Thus, (X, Y) in $An_{\mathcal{P}}$.

($\Rightarrow$) We will first prove that if there is a directed path from X to Y in all MAGs in $[\mathcal{P}]$, and there is no such directed path in $\mathcal{P}$ then there exist at least two potentially directed paths (and thus, two uncovered potentially directed paths) in $\mathcal{P}$:

If there is a directed path from X to Y in every MAG in $[\mathcal{P}]$, then there is a p.d. path from X to Y in $\mathcal{P}$. By Lemma B.1 in [15] there is an uncovered p.d. path from X to Y in $\mathcal{P}$. Let $p_1 = \langle X, U, U_2, \dots U_n, Y\rangle$, be such a path of length n. We want to show that another uncovered p.d. path $p_2 = \langle X, V, V_2, \dots V_m, Y\rangle$ of length m with $U \neq V$ must also exist. Assume there is no such path. If for every MAG in $[\mathcal{P}]$, $X \rightarrow U$, then p_1 is a directed path in $\mathcal{P}$ (since every triple on the path is a definite non collider), and $X \in An_{\mathcal{P}}(Y)$, which is a contradiction. Thus, $X \circ\!\!\rightarrow\!\!\star U$ in $\mathcal{P}$, and there exists a MAG in $[\mathcal{P}]$ where $X \leftarrow\!\!\star U$ and p_1 is not a directed path. If p_1, is the only p.d. path from X to Y in $\mathcal{P}$, then $X \dashrightarrow Y \notin [\mathcal{P}]$, which is a contradiction. Thus, $\exists p_2 = \langle X, V, V_2, \dots V_m, Y\rangle$ with $U \neq V$.

Next we show that $\langle V, X, U\rangle$ form an unshielded definite non-collider in $\mathcal{P}$:
$\langle V, X, U\rangle$ is not a collider in $\mathcal{P}$ (trivially since p_1 and p_2 are potentially directed paths). We must also show that U, V are not adjacent. If $U \star\!\!\rightarrow\!\!\star V$ in , then there exists a MAG in $[\mathcal{P}]$ such that $U \rightarrow X$ in $\mathcal{P}$, which is inconsistent since p_1 is a p.d. path. ■

Apart from (positive) causal relations, a PAG can also have negative and ambiguous causal relations. X and Y share a negative causal relationship in $\mathcal{P}$ if X cannot be a cause of Y in $\mathcal{P}$. This happens if $(X, Y) \notin An_{\mathcal{P}}$, and there can be no directed path from X to Y in $\mathcal{P}$ (i.e. $X \leftarrow\!\!\star Y$ in $\mathcal{P}$, or there is no potentially directed path from X to Y in $\mathcal{P}$). Naturally, this does not mean that Y causes X. An ambiguous causal relationship occurs when a relationship is neither positive nor negative.

We use the notation $NAn_{\mathcal{P}}$ to denote the set of negative causal relationships in a PAG $\mathcal{P}$. The conditions mentioned above for membership in $NAn_{\mathcal{P}}$ can easily be tested in $\mathcal{P}$: To rule out the existence of a possible directed path, only uncovered possibly directed paths need to be checked [15]. Notice that the set of ancestral and non-ancestral relations in a PAG are by no means complementary, since ambiguous relations also exist.

We are interested in constraint-based algorithms' consistency to marginal ancestral sets: Let $\mathcal{D}$ be a data set over a set of possibly confounded variables $\mathbf{V}$, and let $\mathcal{G}$ be the PAG output of a sound and complete constraint-based algorithm. $\mathcal{P}$ defines an ancestral set $An_{\mathcal{P}}$ and a non-ancestral set $NAn_{\mathcal{P}}$. Also, let $\mathcal{P}[_{\mathbf{L}}$, $\mathbf{L} \subset \mathbf{V}$, be the marginal PAG obtained using the same algorithm (with the same hyperparameters) on the restriction of data set $\mathcal{D}$ on variables in $\mathbf{L}$. Each marginal PAG $\mathcal{P}[_{\mathbf{L}}$ defines a marginal ancestral set $An_{\mathcal{P}[_{\mathbf{L}}}$ and a marginal negative ancestral set $NAn_{\mathcal{P}[_{\mathbf{L}}}$.

Assuming no statistical errors, any marginal ancestral set is a subset of the ancestral set. Thus, there is no pair (X, Y) present in any $An_{\mathcal{P}}[_{\mathbf{L}}$ that is not present in $An_{\mathcal{P}}$. On the contrary, some ancestral relationships that can be identified in the full data set may not be identifiable in a marginal, due to (a) the fact that some variables are not included in the marginal and (b) the loss of information from excluding variables. Therefore, members of $An_{\mathcal{P}}$ are possibly not present in some $An_{\mathcal{P}[_{\mathbf{L}}}$. Similarly, any marginal non-ancestral set is a subset of the non-ancestral set, while some

[1] Although theoretically possible, the algorithm assumes that the input PAG and separating sets are correct, that is, that the dependencies and independencies encoded in the PAG are the ones implied by the separating sets. In practice however, as suggested by anecdotal experiments, FCI and its variants rarely produce such output.

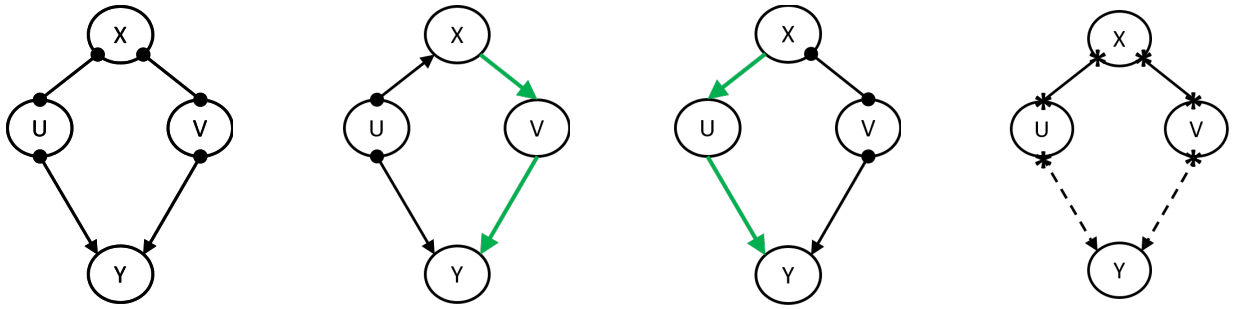


Figure 1: (a) Example of a PAG in which each Markov equivalent MAG contains a directed path from X to Y. (b,c) Orienting the edge between U and X as $U \star\!\!\rightarrow X$ creates the directed path $X \rightarrow V \rightarrow Y$, while orienting it as $U \leftarrow X$ creates the directed path $X \rightarrow U \rightarrow Y$. (d) The general pattern as described in Theorem 3.1. Dashed edges correspond to potentially directed paths.

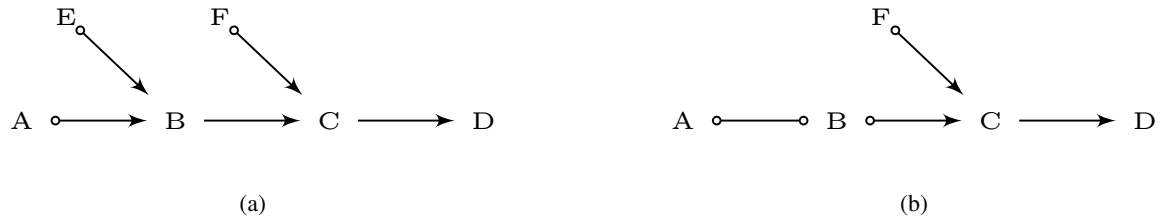


Figure 2: An example of ancestral and non-ancestral sets for a PAG $\mathcal{P}$ and a marginal PAG $\mathcal{P}[_E$.
Ancestral and non-ancestral sets: $An_{\mathcal{P}} = \{(B, \{C, D\}), (C, D)\}$,
$NAn_{\mathcal{P}} = \{(A, \{E, F\}), (B, \{A, E, F\}), (C, \{A, B, E, F\}), (D, \{A, B, C, E, F\}), (E, \{A, F\}), (F, \{A, B, E\})\}$.
The corresponding marginal sets are $An_{\mathcal{P}[_E} = \{(C, D)\}$,
$NAn_{\mathcal{P}[_E} = \{(A, F), (B, F), (C, \{A, B, F\}), (D, \{A, B, C, F\}), (F, \{A, B\})\}$

members of $NAn_{\mathcal{P}}$ may not be present in some $NAn_{\mathcal{P}[_{\mathbf{L}}}$.

In addition, the PAGs should not encode any conflicting causal information: Members of $An_{\mathcal{P}}$ cannot also be members of $NAn_{\mathcal{P}[_{\mathbf{L}}}$, and members of $NAn_{\mathcal{P}}$ cannot also be members of $An_{\mathcal{P}[_{\mathbf{L}}}$.

For finite samples, however, statistical errors propagate in both the skeleton identification and the orientation steps of constraint-based algorithms, and can result in conflicting orientations. Since the algorithms are run on marginal versions of the same data set, conflicts in the marginal ancestral sets can be viewed as a measure of robustness of the algorithm to statistical errors. We are therefore interested in comparing the *marginal causal consistency* of different FCI variations. Marginal consistency can also been as a measure of the sensitivity of an algorithm to the specific choice of observed variables.

For a given algorithm, we are interested in how often positive and negative ancestral relationships entailed by the marginal outputs $\mathcal{P}[_{\mathbf{L}}$ agree or disagree with the output $\mathcal{P}$ over the full set of variables. This information is entailed in the confusion matrix described in Table 1.

Some remarks:

		Marginal PAG $\mathcal{P}[_{\mathbf{L}}$	
		Ancestral	**Non ancestral**
PAG $\mathcal{P}$	**Ancestral**	p	c
	Non ancestral	d	n
	Ambiguous	e	f
		$\lvert An_{\mathcal{P}[_{\mathbf{L}}}\rvert$	$\lvert NAn_{\mathcal{P}[_{\mathbf{L}}}\rvert$

Table 1: Confusion matrix for (non) ancestral relationships for a marginal of $\mathcal{P}$.

		Marginal PAG $\mathcal{P}[_{\mathbf{L}}$	
		Ancestral	**Non ancestral**
PAG $\mathcal{P}$	**Ancestral**	2	0
	Non ancestral	0	11
	Ambiguous	0	0
		2	11

Table 2: Confusion matrix for the PAGs in Figure 2.

- For perfect statistical knowledge, $c = d = e = f = 0$, $p = |An_{\mathcal{P}[_{\mathbf{L}}}|$, $n = |NAn_{\mathcal{P}[_{\mathbf{L}}}|$. Notice that ancestral and non-ancestral relationships identified in a marginal PAG are not expected to be ambiguous in the PAG over the complete set of variables $\mathbf{V}$.

- The sum $p + d + c + n + e + f$ is different (smaller) than the number of possible causal relationships in the marginal data set, since $\mathcal{P}[_{\mathbf{L}}$ also has ambiguous edges. We do not take these edges into account here, because they are not an indication of consistency or inconsistency of the marginal. Edges that are (non) ancestral in $\mathcal{P}$ can often be ambiguous in $\mathcal{P}[_{\mathbf{L}}$, even if the endpoints are present in the marginal.

An example of (non) ancestral relations in a PAG and a corresponding marginal PAG is shown in Figure 2. The corresponding confusion matrix is shown in Table 2 (assuming perfect statistical knowledge).

Notice however, that the confusion matrix in Table 1 only measures the robustness of an algorithm in terms of *causal* predictions. Other characteristics and uncertainties of FCI outputs are not taken into account.

3.1 Marginal consistency of FCI variations in simulated data.

We used simulated data to examine the internal consistency of the FCI algorithm. We used the pcalg package [8] to simulate random DAGs with 20 variables. We tried two different graph densities, 0.1 and 0.2 (corresponding to 1.9 and 3.8 neighbors per variable on average, respectively). We use linear Gaussian parametrization, with coefficient of each model sampled using the default parameters in the pcalg package. For each graph density we generated 50 DAGs and for each such network we simulated data with 1000 samples.

We used different variations of the FCI algorithm to retrieve a causal network with significance threshold $\alpha = 0.05$ and unconstrained maximum conditioning set. We also created 100 randomly selected marginal data sets of size 18 and 15 for each DAG, and ran all FCI variations with the same parameters.

The variations of the FCI algorithm we used are the following: order independent FCI (iFCI), FC, FCI with majority rule (mFCI). We did not include the conservative FCI, as the existence of ambiguous triples results in outputs that are not complete PAGs. Hence, Theorem 3.1 can not be applied. Instead, we included the majority rule FCI, which is inspired by conservative FCI, but in which the triplet's orientation is dictated by a majority vote on the corresponding conditional independence tests. To guarantee that the output is a valid PAG, ambiguous triplets (occurring in a tie vote) are marked as definite non-colliders.

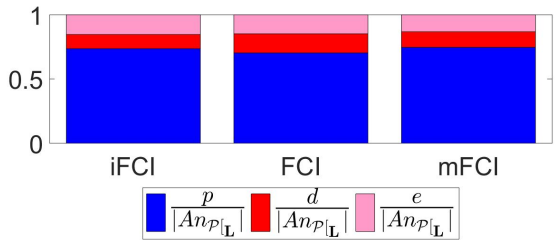
Due to statistical errors, the output of the FCI algorithm is also not necessarily a valid PAG. A very common problem is the creation of cycles or almost cycles. To tackle this problem, we added the option to aggressively prevent cycles, as implemented in TETRAD [13]. This functionality is applied in the phase of orienting edges, where every attempted orientation is checked for creating an (almost) cycle. If that is the case, then that specific orientation is not performed, and the orientation rules move on to the next possible orientation. We have to note that we do not use the orientation rules that aim at recovering undirected edges (selection bias).

The results of the experiments are shown in Figures 3 and 4. The ratios were computed by summing over all numerators and dividing by the sum of all denominators (for example, for $\frac{p}{|An_{\mathcal{P}[_{\mathbf{L}}}|}$ we summed over all correctly identified ancestral relations p and divided by the total number of predicted ancestral relations $|An_{\mathcal{P}[_{\mathbf{L}}}|$ over all marginals and datasets). This guarantees that each bar sums to one and avoids divisions by zero, in case no relation is predicted.

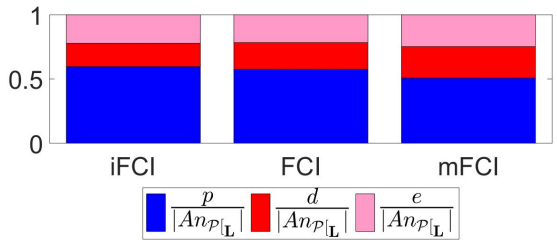
For all algorithms, the consistency drops for both denser networks and smaller marginals. For networks with density 0.1 all algorithms have more than 50% consistent predictions for both 18 and 15-variable marginals. For denser networks however, the performance of iFCI and FCI drops below 0.2. mFCI has the largest ratio of consistent relationships, and retains a consistency ratio of 0.60 for 18-variable marginals. However, its performance drops to 0.38 for 15-variable marginals. For all algorithms, the majority of causal relationships are found non-ancestral in $\mathcal{P}$, and a small ratio is found ambiguous. It is worth noting, however, that the algorithms typically output very few positive causal relationships.

Non-ancestral causal relationships on the other hand are much more consistent, as shown in Figures 5 and 6. The majority of non-ancestral relationships are consistent (blue bars, $\frac{n}{|An_{\mathcal{P}[_{\mathbf{L}}}|}$). Overall, mFCI has the highest ratio of consistent negative relationships.

Overall, results show that (a) the performance of constraint-based algorithms heavily depends on the graph density, particularly for algorithms that are less conservative, and thus more sensitive to error propagation and (b) The causal predictions of the algorithms are sensitive to marginalization. Even for mFCI, removing two out of 20 variables results in 30-40% relations that are not validated in the marginal data set.

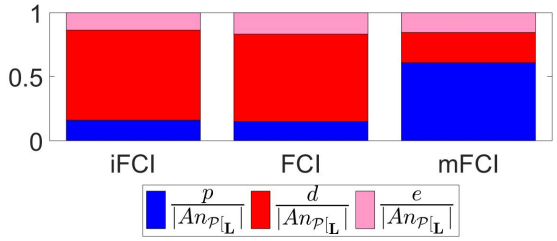


(a)

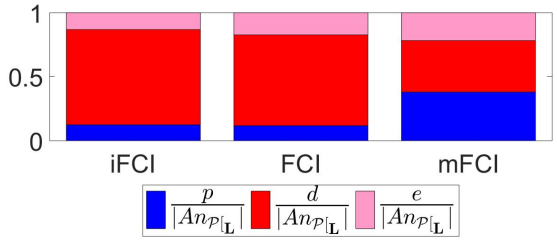


(b)

Figure 3: Barplots showing the ratios of positive relations in (a) 18-variable marginals, (b) 15-variable marginals in the graphs of density 0.1. Blue: Consistent ancestral relationships. Red: Inconsistent ancestral relationships (found non-ancestral in $\mathcal{P}$). Pink: Inconsistent ancestral relationships (found ambiguous in $\mathcal{P}$).

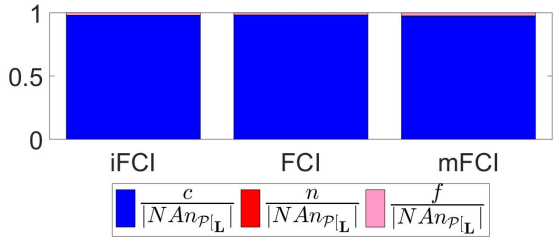


(a)

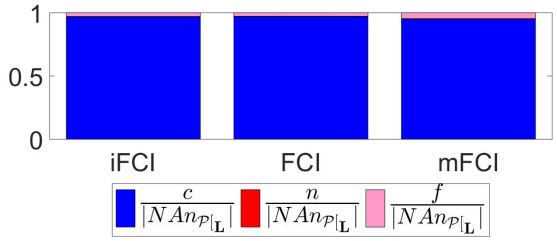


(b)

Figure 4: Barplots showing the ratios of positive relations in (a) 18-variable marginals, (b) 15-variable marginals in the graphs of density 0.2. Blue: Consistent ancestral relationships. Red: Inconsistent ancestral relationships (found non-ancestral in $\mathcal{P}$). Pink: Inconsistent ancestral relationships (found ambiguous in $\mathcal{P}$).



(a)



(b)

Figure 5: Barplots showing the ratios of negative relations in (a) 18-variable marginals, (b) 15-variable marginals in the graphs of density 0.1. Blue: Consistent non-ancestral relationships. Red: Inconsistent non-ancestral relationships (found ancestral in $\mathcal{P}$). Pink: Inconsistent non-ancestral relationships (found ambiguous in $\mathcal{P}$). The majority of non-causal predictions are consistent (blue).

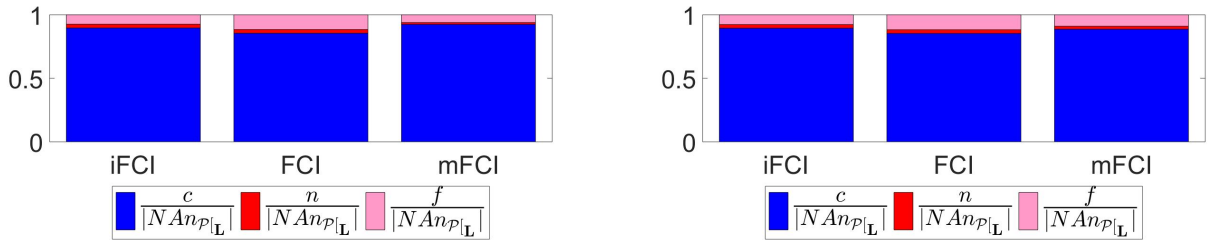
4 RANKING CAUSAL RELATIONSHIPS BASED ON MARGINAL CAUSAL CONSISTENCY

Calculating the marginal ancestral sets indicates a way of ranking pairwise causal relationships according to their frequency of appearance in $An_{\mathcal{P}[\mathbf{L}}$ for different marginals. The idea is that causal relationships that frequently appear in marginal PAGs will tend to be true more often, even if they are not consistent with the output of the algorithm on the whole data set.

4.1 Related Work

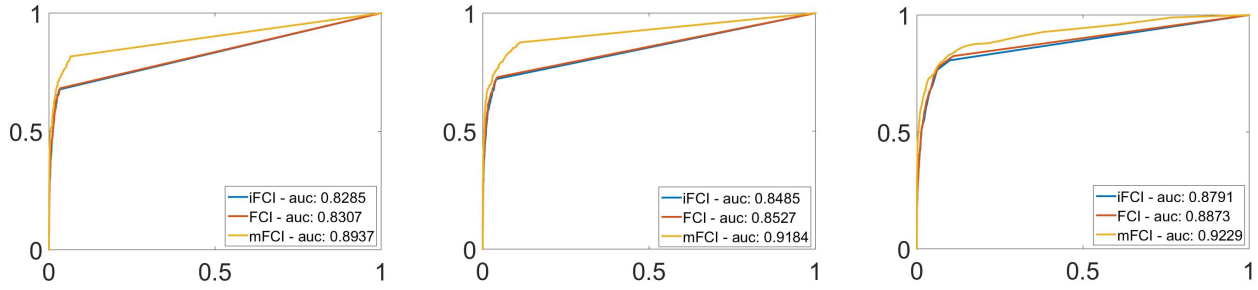
Alternative approaches for ranking causal ancestry relations can be categorized into (a) Bayesian model averaging methods and (b) resampling-based methods.

Bayesian model averaging methods compute the posterior



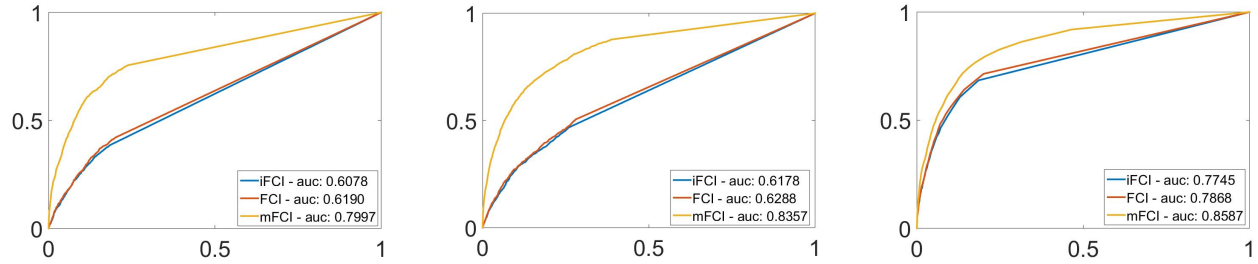
(a) (b)

Figure 6: Barplots showing the ratios of negative relations in (a) 18-variable marginals, (b) 15-variable marginals in the graphs of density 0.2. Blue: Consistent non-ancestral relationships. Red: Inconsistent ancestral relationships (found ancestral in $\mathcal{P}$). Pink: Inconsistent non-ancestral relationships (found ambiguous in $\mathcal{P}$). The majority of non-causal predictions are consistent (blue).



(a) (b) (c)

Figure 7: ROC curves of the variations of the FCI algorithm in graphs with originally 20 variables and density 0.1. (a) 18-variable marginals (b) 15-variable marginals (c) bootstrapping.



(a) (b) (c)

Figure 8: ROC curves of the variations of the FCI algorithm in graphs with originally 20 variables and density 0.2. (a) 18-variable marginals (b) 15-variable marginals (c) bootstrapping.

probability of network features by averaging over network structures. This can be either approximated using MCMC [6] or by using exact methods [10, 2]. Apart from their high computational cost, such methods are not very general and are only applicable to cases where network scores can be computed. Although various scores exist for Bayesian network structures [7], scores for MAGs have only recently been explored [12] and require more expensive fitting procedures [4].

Resampling-based methods repeatedly apply a learning method on resampled datasets and estimate the confidence of network features as the proportion of induced networks in which they appeared. Such methods include parametric and non-parametric bootstraps [5], as well as stability selection [9]. The main advantage is that they are general and thus also applicable in the case of PAGs.

In this section we compare our approach with the non-

parametric bootstrap by [5].

4.2 Experiments

We produced rankings for all possible pairs of variables according to (a) the frequency of appearance in the corresponding marginal ancestral sets $An_{\mathcal{P}[\mathbf{L}}$ and (b) the frequency of appearance in $An_{\mathcal{P}}$ for FCI outputs on different bootstrap samples of the initial data set D. Since we only use 100 different marginals per iteration, for each pair (X, Y), the frequency of appearance in $An_{\mathcal{P}[\mathbf{L}}$ was divided by the number of data sets in which X and Y are both present (i.e. we excluded from the calculation marginals in which the causal ancestry could not be found). For every pair of variables the true label is 1 if the relationship is ancestral in the DAG the data was sampled from, and 0 otherwise.

Based on these ranking and the status of the relationship in the ground truth network (used to simulate the data), we calculated ROC curves and the corresponding AUCs. Figures 7(a,b) and 8(a,b) show the performance of the algorithms using marginals of 15 and 18 variables for densities 0.1 and 0.2 respectively. For all settings, ROC curves are significantly better than random guessing. Again, the results are better for sparser networks, where AUC ranges from 82.25-91.84%. The AUCs drop for denser networks, ranging from 60.78-83.57%. mFCI has the highest AUCs for all settings.

We compared our method against bootstrapping. For each DAG, 100 data sets with 1000 samples were resampled with replacement from the original data set. FCI was ran on the new data sets, each time calculating the ancestral set $An_{\mathcal{P}}$ on the output PAG. Each order pair of variables was then ranked according to the frequency of appearance in the ancestral sets. The results are shown in figures 7(c) and 8(c), where we can see that bootstrapping outperforms the marginal-based ranking in all cases. Again, the majority rule FCI outperforms the rest of the algorithms.

5 DISCUSSION

We defined and examined the problem of marginal causal consistency in constraint-based causal discovery. We presented a way to identify all invariant causal relationships entailed in a complete PAG.

We examined how well causal and non-causal relationships predicted by three different variations of FCI are preserved when you marginalize out variables. Results indicate that constraint-based learning methods for causal networks are sensitive to the selected marginal, particularly for dense networks.

The results are important because in most real-life applications researchers may have limited knowledge on the possible unmeasured variables. It is also possible that confidence metrics computed based on marginals could be more helpful in situations where you have "outlier" variables that do not satisfy the algorithm's assumptions (e.g. they create cycles, do not satisfy the distributional assumptions of conditional independence tests etc). In such cases, it is possible that taking random marginals could improve the algorithms' performance (similar to the way bootstrapping is beneficial when you have outlier samples).

We must also point out that PAGs encode much richer information than the causal ancestry relations examined here. Exploring different types of marginal consistency is also an area of interest.

Acknowledgments

We would like to thank the anonymous reviewers for their comments, particularly reviewer 2 who helped us identify a problem in the simulations in the submitted version of this paper. This work was funded by European Research Council (ERC) and is part of the CAUSALPATH - Next Generation Causal Analysis project, No 617393.

References

[1] G. Borboudakis and I. Tsamardinos. Incorporating causal prior knowledge as path-constraints in Bayesian networks and maximal ancestral graphs. *Proceedings of the 29th International Conference on Machine Learning*, pages 1799–1806, 2012.

[2] Y. Chen, L. Meng, and J. Tian. Exact bayesian learning of ancestor relations in bayesian networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

[3] D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[4] M. Drton, M. Eichler, and T. S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10:2329–2348, 2009.

[5] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[6] N. Friedman and D. Koller. Being Bayesian about network structure. In *Proceedings of the Sixteenth Conference Annual Conference on Uncertainty in Artificial Intelligence*, 2000.

[7] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[8] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

[9] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:417–473, 2010.

[10] P. Parviainen and M. Koivisto. Ancestor relations in the presence of unobserved variables. In *Machine Learning and Knowledge Discovery in Databases*, volume 6912 of *Lecture Notes in Computer Science*, pages 581–596. Springer, 2011.

[11] J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 2006.

[12] T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

[13] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

[14] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[15] J. Zhang. 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, 2008.