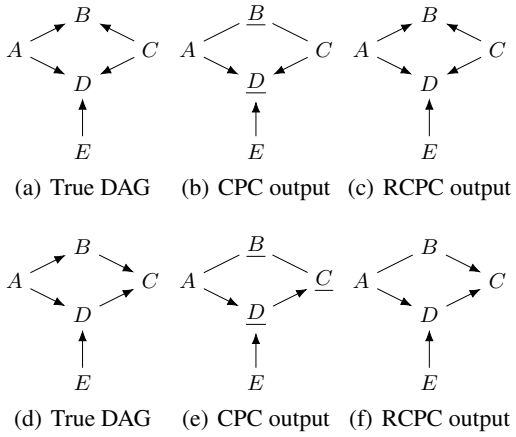
RCPC: A Sound Causal Discovery Algorithm under Orientation Unfaithfulness

Kenneth Lee[1] **Murat Kocaoglu**[1]

[1]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA

Abstract

In causal discovery, the constraint-based approaches often rely on an assumption known as faithfulness/stability, only the variables that are d-separated in a directed acyclic graph will be statistically independent. This assumption can be partitioned into two subconditions: orientation faithfulness and adjacency faithfulness. Under adjacency faithfulness, a sound algorithm known as CPC, a conservative version of PC algorithm, has been developed and is conjectured to be complete. In this work, we show that the CPC algorithm is not complete and propose two new sound orientation rules as part of a sound causal discovery algorithm called revised CPC (RCPC) under orientation unfaithfulness.



(a) True DAG (b) CPC output (c) RCPC output

(d) True DAG (e) CPC output (f) RCPC output

Figure 1: Two examples to showcase how CPC is not complete. Top: example 1 with (a) true DAG, (b) the CPC output (c) RCPC output; Bottom: example 2 with (d) true DAG, (e) the CPC output (f) RCPC output

1 INTRODUCTION

In constraint-based causal discovery, the learning of causal graphs relies on a set of constraints that the graph structure imposes on all probability distributions compatible with the graph. Some of these constraints are found by performing a series of conditional independence tests in a large sample limit to rule out impossible graph structures. One of the variants of PC algorithm, known as the conservative PC algorithm (CPC) relaxes one of the common assumptions called faithfulness [Ramsey et al., 2012], which can often be violated in practice due to finite sample and a series of work have been dedicated to study and work around this issue [Robins et al., 2003, Koivisto and Sood, 2004, Shimizu et al., 2006, Zhang and Spirtes, 2012, Uhler et al., 2013, Spirtes and Zhang, 2014, Zhang et al., 2017, Solus et al., 2017, Raskutti and Uhler, 2018, Lin and Zhang, 2020, Bernstein et al., 2020, Lu et al., 2021, Ghassami et al., 2020, Marx et al., 2021, Ng et al., 2021]. CPC has been conjectured to be complete in the sense that it can recover up to an equivalence class of DAGs without missing any edges or orientations in the true DAG that can potentially be learned with a given set of constraints.

1.1 CPC ALGORITHM IS NOT COMPLETE

Assuming adjacency faithfulness, consider the example 1 in Figure 1 where we have a probability distribution p over G such that p is orientation unfaithful to G where $(A \perp\!\!\!\perp C)_p, (A \perp\!\!\!\perp E)_p, (A \not\perp\!\!\!\perp C \,|B)_p, (B \perp\!\!\!\perp D \,|A, C)_p, (A \perp\!\!\!\perp C \,|B, D)_p$. In this case, orientation-faithfulness does not hold because of the triples $\langle A, B, D\rangle$ and $\langle A, C, D\rangle$. Given the correct conditional independence oracle, CPC will orient the triple $\langle A, B, C\rangle$ as $A - \underline{B} - C$, as shown by Figure 1(b), due to the fact that $(A \perp\!\!\!\perp C)_p$ and $(A \perp\!\!\!\perp C \,|B, D)_p$ do not satisfy the conditions in steps 3(A) and 3(B) of the algorithm and Meek rules are only applied to triples that are not marked as unfaithful. However, under causal Markov

assumption with $(A \not\perp\!\!\!\perp C \,|B\,)_p$, we see that it is impossible to orient $\langle A, B, C\rangle$ as a non-collider. Therefore, we should orient $\langle A, B, C\rangle$ as $A \to B \leftarrow C$, which yields the result in Figure 1(c). We remove the underline because the ambiguity concerning whether the triple is a collider or a non-collider has been dissolved. This motivates us to develop the RCPC algorithm.

2 RCPC ALGORITHM

Our modification will add the following steps after step 4 of the CPC algorithm. We leave the relevant definitions in Appendix A.

We will recursively apply **R5** and **R6** until there is no more edges that can be oriented by them.

R5: Let G be the resulting graph from step 4 or step 6 (after going through step 5 and step 6 for the first time) and H be the set that contains all subsets of A's potential parents and of C's potential parents, for each unshielded triple $\langle A, B, C\rangle$ that has been marked unfaithful in G with distribution p:

- If it has been oriented as $A \to \underline{B} \leftarrow C$ by other triples, mark $(A \perp\!\!\!\perp C \,|W\,)_p$ as NM (Non-Markov) statement and unmark $A \to \underline{B} \leftarrow C$ as $A \to B \leftarrow C$, for any W that contains B. Also, if all cancelled paths from A to C relative to B are along all the d-connecting paths from X to Y relative to J, then we also mark $(X \perp\!\!\!\perp Y \,|J\,)_p$ as NM (Non-Markov) statement for any J that contains B, where $J \in Q \setminus \{B\}$, where Q be the set that contains all subsets of X's potential parents and of Y's potential parents and $X, Y \in V$.
- If it has been oriented as $A \leftarrow \underline{B} \to C$ or $A \leftarrow \underline{B} \leftarrow C$ or $A \to \underline{B} \to C$ or $A \leftarrow \underline{B} - C$ or $A - \underline{B} \to C$ by other triples, mark $(A \perp\!\!\!\perp C \,|S\,)_p$ as NM statement and unmark the triple from being unfaithful, for any S that does not contain B, where $S \in H$. Also, if all cancelled paths from A to C relative to $\{\emptyset\}$ is along all the d-connecting paths from X to Y relative to D, then we also mark $(X \perp\!\!\!\perp Y \,|D\,)_p$ as NM statement for any D that does not contain B, where $D \in Q$, $X, Y \in V$.

Then, without considering any conditional independence relation that has been marked as NM statements, for each unshielded triple $\langle A, T, C\rangle$ in G with a distribution p, check all subsets of A's potential parents and of C's potential parents:

(a) If T is NOT in any such set conditional on which A and C are independent, orient $A - \underline{T} - C$ as $A \to T \leftarrow C$

(b) If T is in all such sets conditional on which A and C are independent, orient $A - \underline{T} - C$ as $A - T - C$

(c) otherwise, keep the triple as "unfaithful" by underlining the triple

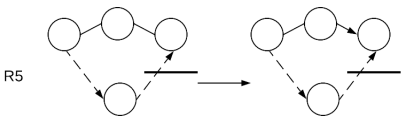


Figure 2: This is an additional orientation rule in R6 to orient an edge for any length four or above cycle due to acyclicity and the result of unshielded non-colliders shown by the solid lines. The underline represents the unshielded triple of the form $A - \underline{B} \leftarrow C$ being marked as unfaithful. The dash line indicates a directed path. We keep the underline so that R5 of RCPC may utilize it for adding NM statements.

R6: Recursively apply R1, R3, and R4 of Meek rules Meek [2013] to unshielded non-colliders that are not marked as unfaithful in G except that R2 can be applied to any edges. Then, orient edges by successive application of an additional rule as shown by Figure 2 with the following conditions on the subgraphs: the triple should not be unmarked so that R5 can utilize it for adding NM statements.

To illustrate the usage of the orientation rule R6, consider the example shown by Figure 1(d)- 1(f) where we have a probability distribution p over G such that f is orientation unfaithful to G yielding $(E \perp\!\!\!\perp A)_p, (E \perp\!\!\!\perp B)_p, (A \perp\!\!\!\perp C)_p, (A \perp\!\!\!\perp C|B, D)_p, (B \perp\!\!\!\perp D|A)_p, (B \perp\!\!\!\perp D|C, A)_p, (A \not\perp\!\!\!\perp C|B)_p, (A \not\perp\!\!\!\perp C|D)_p$. In this case, the CPC algorithm will output $A - \underline{B} - \underline{C}; A \to \underline{D} \to \underline{C}; E \to \underline{D}$ as illustrated in Figure 1(e). However, we see that $A - \underline{B} - \underline{C}$ cannot be a collider due to causal Markov assumption so that we can unmark $\langle A, B, C\rangle$ from being unfaithful. Note that RCPC will also orient $A - \underline{B} - \underline{C}$ as $A - B - \underline{C}$ since $(A \perp\!\!\!\perp C)_p$ has been marked as NM statement by checking the triple $\langle A, D, C\rangle$ with R5. Then, due to acyclity, we can then orient $A - B - \underline{C}$ as $A - B \to C$ by applying an additional rule in R6 to get the result shown by Figure 1(f). The unfaithful mark can be removed as we have resolved the ambiguity of being a collider vs a non-collider for those triples.

Theorem 2.1. (Soundness of RCPC) Under the causal Markov and Adjacency-Faithfulness assumptions, the RCPC algorithm is correct in the sense that given a perfect conditional independence oracle, the algorithm returns an extended pattern that represents the true causal DAG.

3 CONCLUSION

In conclusion, we demonstrate how CPC algorithm is not complete and provide a sound causal discovery algorithm by leveraging the causal Markov assumption under adjacency faithfulness. We further show that RCPC can have advantages over CPC in terms of recovering an equivalence class that contains the underlying causal graph in the presence of orientation unfaithfulness.

References

Daniel Bernstein, Basil Saeed, Chandler Squires, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4098–4108. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/bernstein20a.html.

Amiremad Ghassami, Alan Yang, Negar Kiyavash, and Kun Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3494–3504. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ghassami20a.html.

Mikko Koivisto and Kismat Sood. Exact bayesian structure discovery in bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004.

Hanti Lin and Jiji Zhang. On learning causal structures from non-experimental data without any faithfulness assumption. In *Algorithmic Learning Theory*, pages 554–582. PMLR, 2020.

Ni Y Lu, Kun Zhang, and Changhe Yuan. Improving causal discovery by optimal bayesian network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8741–8748, 2021.

Alexander Marx, Arthur Gretton, and Joris M Mooij. A weaker faithfulness assumption based on triple interactions. In *Uncertainty in Artificial Intelligence*, pages 451–460. PMLR, 2021.

Christopher Meek. Causal inference and causal explanation with background knowledge, 2013.

Ignavier Ng, Yujia Zheng, Jiji Zhang, and Kun Zhang. Reliable causal discovery with improved exact search and weaker assumptions. *Advances in Neural Information Processing Systems*, 34:20308–20320, 2021.

Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference, 2012.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

James M Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms, 2017.

Peter Spirtes and Jiji Zhang. A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption. *Statistical Science*, pages 662–678, 2014.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Jiji Zhang and Peter L Spirtes. Strong faithfulness and uniform consistency in causal inference, 2012.

Jiji Zhang, Wolfgang Mayer, et al. Weakening faithfulness: some heuristic causal discovery algorithms. *International journal of data science and analytics*, 3(2):93–104, 2017.

RCPC: A Sound Causal Discovery Algorithm under Orientation Unfaithfulness (Supplementary Material)

Kenneth Lee[1] **Murat Kocaoglu**[1]

[1]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA

A SOURCES OF UNFAITHFULNESS

Here, we introduce some conditions that occur under unfaithfulness and are necessary to understand the workings of our new algorithm **RCPC**.

Definition A.1. (Cancelled paths) In a DAG $G = (V, E)$ with any unfaithful distribution p compatible with G, we say the active paths **q** between a set of variables $\mathbf{X}$ and another set of variables $\mathbf{Y}$ are *cancelled* relative to a set of vertices $\mathbf{Z} \subseteq V, (\mathbf{X}, \mathbf{Y} \not\subseteq \mathbf{Z})$ if $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \,|\mathbf{Z})_G$ and $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{Z})_p$

For instance, in example 1 as shown by Figure 1(a)-1(c), the path $A \to B \leftarrow C$ cancels the path $A \to D \leftarrow C$ and we call those two paths together cancelled paths. Note that it is also possible to have $A \to B \leftarrow C$ being a cancelled path alone, we call such path a *self-cancelling* path.

Definition A.2. (UF-connecting and separation) In a DAG $G = (V, E)$, a path q between vertices $\mathbf{X}$ and $\mathbf{Y}$ is *UF-connecting* relative to a set of vertices $\mathbf{Z} \subseteq V (X, Y \not\subseteq \mathbf{Z})$ if every non-collider on q is not a member of $\mathbf{Z}$; (ii) every collider on q is an ancestor of some members of $\mathbf{Z}$; and (iii) any subpath of q is not a cancelled path relative to $\mathbf{Z}$.any subpath of q is not a cancelled path relative to $\mathbf{Z}$. Two sets of variables $\mathbf{X}$ and $\mathbf{Y}$ are said to be *UF-separated* by $\mathbf{Z}$ if there is no UF-connecting path between any member of $\mathbf{X}$ and any member of $\mathbf{Y}$ relative to $\mathbf{Z}$.

Theorem A.1. (Probabilistic implication of UF-separation) If sets $\mathbf{X}$ and $\mathbf{Y}$ are UF-separated by $\mathbf{Z}$ in a DAG G, then $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in every unfaithful distribution compatible with G. Conversely, if $\mathbf{X}$ and $\mathbf{Y}$ are not UF-separated by $\mathbf{Z}$ in a DAG G, then $\mathbf{X}$ and $\mathbf{Y}$ are dependent conditional on $\mathbf{Z}$ in at least one unfaithful distribution compatible with G.

Theorem A.2. (Propagation of cancelled paths) In a DAG $G = (V, E)$, for disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{U}, \mathbf{T} \subseteq V$, if there exists cancelled path(s) from $\mathbf{X}$ to $\mathbf{Y}$ relative to $\mathbf{Z}$ along all the d-connecting paths from $\mathbf{U}$ to $\mathbf{T}$ relative to $\mathbf{Z}$ or one of following conditions hold

- $(\mathbf{U} \perp\!\!\!\perp \mathbf{T} \,|\mathbf{Z}, \mathbf{X})_p$ and $((\mathbf{U} \perp\!\!\!\perp \mathbf{X} \,|\mathbf{Z})_p \vee (\mathbf{T} \perp\!\!\!\perp \mathbf{X} \,|\mathbf{Z})_p)$
- $(\mathbf{U} \perp\!\!\!\perp \mathbf{T} \,|\mathbf{Z}, \mathbf{Y})_p$ and $((\mathbf{U} \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{Z})_p \vee (\mathbf{T} \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{Z})_p)$

, then $\mathbf{Z}$ UF-separates $\mathbf{U}$ and $\mathbf{T}$.

B PROOF OF THEOREM 2.1

Proof. Suppose the true causal graph is G, and all conditional independence judgments are correct. The correctness of Step 1, 2, 3(A)-(B), and 4 of the RCPC algorithm follows from the Theorem 1 in Ramsey et al. [2012]. Now consider the added steps R5, and R6. At R5(a), for an unshielded triple $\langle A, T, C \rangle$ that has been marked unfaithful, for the sake of contradiction, suppose $\langle A, T, C \rangle$ is a non-collider. By causal Markov assumptions, T is in at least one subset of all subsets of

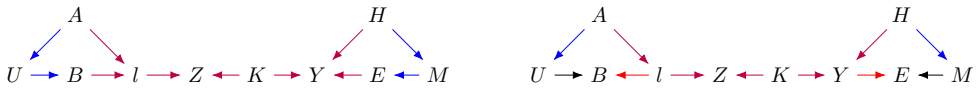


Figure 3: an example illustrates the idea in Theorem A.2. Suppose we have a distribution p such that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, Z)_p$, where $\mathbf{X} = \{A, B\}$ and $\mathbf{Y} = \{H, E\}$. Then, the paths from $\mathbf{X}$ to $\mathbf{Y}$ are *cancelled* relative to Z such that Z UF-separates U and M in D_1 (left) since the cancelled paths from $\mathbf{X}$ to $\mathbf{J}$ relative to Z are all along with d-connecting path from U to M relative to Z, whereas the cancelled paths from $\mathbf{X}$ to $\mathbf{Y}$ relative to Z *are not* all along with d-connecting path from U to M relative to Z in D_2 (right). **Blue**: the d-connecting paths between U and M relative to Z that does not overlap with the cancelled paths. **Red**: The portion of the cancelled path from A, B to H, E relative to Z that does not overlap with the d-connecting paths from U to M relative to Z. **Purple**: The overlapping portion between the cancelled paths from A, B to H, E relative to Z and the d-connecting paths from U and M relative to Z.

A's potential parents and of C's potential parents conditional on which A and C are independent. However, the conditional independence statements based on these subsets do not belong to the NM statements and for all subsets of A's potential parents and of C's potential parents conditional on which A and C, T is not in any such set after excluding the conditional independence based on the NM statements. Therefore, it is a contradiction. Thus, $\langle A, T, C\rangle$ is a collider. For R5(b), for the sake of contradiction, suppose $\langle A, T, C\rangle$ is a collider, by the causal Markov assumption, there exists a subset of all subsets of A's potential parents and of C's potential parents conditional on which A and C are independent that does not contain T, which gives a contradiction since T is in all such sets and such conditional independence does not belong to the NM statements. However, T is in all such sets conditional on which A and C are independent after excluding the conditional independence in the NM statements, which is a contradiction. Therefore, $\langle A, T, C\rangle$ is a non-collider. At R6, the soundness of the rules R1, R3, and R4 follows the step 4 of the CPC algorithm Ramsey et al. [2012], Meek [2013]. For recursively applying R2 of Meek rules, it follows the acyclicity assumptions. For the correctness of the additional rule in R6, without loss of generality, suppose there exists an unshielded triple $\langle A, T, C\rangle$ oriented as $A - T - \underline{C}$ after R5 and there is another directed path $A \to \ldots \to C$ as illustrated by Figure 2, we know that $A - T - \underline{C}$ is not a collider since it is either being unmarked from being unfaithful in S5 or it is not marked as unfaithful in step 3 and it is impossible to have $A \leftarrow T \leftarrow C$ due to acyclicity. Therefore, $\langle A, T, C\rangle$ can only be oriented as either $A \leftarrow T \to \underline{C}$ or $A \to T \to \underline{C}$. Thus, the triple should be oriented as $A - T \to \underline{C}$. □

C ALGORITHMS

Algorithm 1 CPC Ramsey et al. [2012]

Step 1
Form the complete undirected graph U on the set of variables V
Step 2
Initialize $n = 0$
repeat
For each pair of variables X and Y that are adjacent in (the current) U such that $ADJ(U, X) \setminus \{Y\}$ or $ADJ(U, Y) \setminus \{X\}$ has at least n elements, check through the subsets of $ADJ(U, X) \setminus \{Y\}$ and the subsets of $ADJ(U, Y) \setminus \{X\}$ that have exactly n variables. If a subset S is found conditional on which X and Y are independent, remove the edge between X and Y in U, and record S as $Sepset(X, Y)$;
until for each ordered pair of adjacent variables X and Y, $ADJ(U, X) \setminus \{Y\}$ has less than n elements.
Step 3
Let G be the resulting graph from Step 2. For each unshielded triple $\langle A, B, C\rangle$, check all subsets of A's potential parents and of C's potential parents:
(A): If B is NOT in any such set conditional on which A and C are independent, orient $A - B - C$ as $A \to B \leftarrow C$
(B): if B is in all such sets conditional on which A and C are independent, orient $A - B - C$
(C): otherwise, mark the triple as "unfaithful" by underlining the triple $A - \underline{B} - C$
Step 4
Apply Meek rules Meek [2013] to unshielded non-colliders, not including triples that are marked as unfaithful in G.