Manuscript ID: DGJCI.2017.0020
Title: "The Inflation Technique for Causal Inference with Latent Variables"
Authors: E. Wolfe, R.W. Spekkens, and T. Fritz

Journal of Causal Inference (JCI)
Author's Response to Reviewer/Editor Critique.

We are grateful to the thoughtful comments of the referees, especially in regard to concrete suggestions for making the manuscript more accessible to non-physicists. We have substantially revised the manuscript accordingly. We feel that the manuscript has been appreciably improved by the revisions we have implemented following the referees comments.

We reply to the numbered points of the referees below.

Response to Referee 1

1) *In the definition of compatibility the requirement is that the distribution contains independence relations of the structure and the distribution can contain much more conditional relations. In other words, faithfulness of the distribution is never considered. This can extremely restrict the usefulness of the defined compatibility in algorithmic structure learning. For example, on page 3, it is said that: "The DAGs that are compatible with the given distribution can be determined algorithmically". This is in general seems to be not true and faithfulness is required too. Some comments from the authors in this regard would be appreciated.*

1A) The referee is correct in remarking that this manuscript does not account for faithfulness of the distributions. We have tried to make the scope of this manuscript clear by stating in the introduction *"A special case of both problems is the following decision problem: given a probability distribution and a hypothesis about the causal relations, determine whether the two are compatible: could the given distribution have been generated by the hypothesized causal relations? This is the problem that we focus on."*

1B) The question of finding a causal hypothesis for which the given probability distribution is both compatible and faithful is not the ultimate goal of our work, though our contribution of an algorithm for testing compatibility can be composed with algorithms for testing faithfulness in order to perform the causal inference of the sort advocated by Referee 1. Efficient algorithms for testing faithfulness exist independently of this work. Thus, one way to isolate causal explanation satisfying both compatibility and faithfulness constraints is to start with the set of *all* DAGS over the given observed variables, filter out the subset of DAGS such that all conditional independence relations in the distribution are reflected as d-separation relations in every DAG in that subset. This gives the set of DAGs for which the given distribution is faithful. Then, one can apply the Inflation Technique to each such DAG to rule out those causal explanation not actually compatible with the given distribution.

1C) In summary, while tests for compatibility and tests for faithfulness are both relevant to causal inference, they are distinct conceptual desiderata. As such, we see the uncluttered focus on causal compatibility in this manuscript to be a positive feature, not an oversight.

2) *In the definition of inflation, Definition 2, G' can have significantly less nodes than G, while in the rest of the paper, it seems that it is always useful that the inflated structure has at least one copy of each variable in G. Was there a reason for this specific definition of inflation.*

2A) The referee is correct in noting that the added generality in Definition 2 for allowing fewer nodes in the inflated structure is not exploited in in any of the examples in the paper. We felt that this generality might be important, however. As a simple illustration, imagine that the Triangle structure appears as an ancestral subgraph of some much larger DAG. The inflations of the Triangle structure, such as the Cut inflation, plainly leads to meaningful constraints on all distributions compatible with this hypothetical larger DAG by virtue of constraining the three-observed-variables marginal distributions.

2B) In a separate manuscript [arXiv:1707.06476] an explicit hierarchy of potential inflated structures is described for any DAG; in that hierarchy every inflated structure G' contains more nodes than the original graph G, in line with the referee's intuition.

3) *In equation (6)-(9), it seems that U~U' is redundant.*

3A) Actually no, this condition is required to align the variables across the two sets. As a simple counterexample, consider the two-node graph A→B, and consider the trivial inflation A_1→B_1. Then, AnSubDAG(B) ~ AnSubDAG(A_1, B_1), but nevertheless the singleton set {B} is not copy-index-equivalent to the multivariate set {A_1, B_1}. The requirement U~U' makes it clear that we are only comparing set of variables which are one-to-one copy-index equivalent.

4) *In several places it is mentioned that the inflation technique applies just as well in the case of continuous variables. This is true for the general idea, e.g., by using information theoretic inequalities as presented in one of the examples, but for the general approach provided in Section IV, it is not clear.*

4A) We find the referee's objection valid, and we have walked back the claim of application to continuous variables from complete generality to only special cases. Though information theoretic inequalities do constitute an example, they are not the special cases we had in made. Accordingly, we have now added explicit text to Example 4 in the main text in which we show that an inequality in terms of expectation values applies to both binary and continuous variables. Unlike entropic inequalities, this example is in line with the general technique later presented.

5) *An important missing parts in this work is to provide a clear method or at least some rigorous ideas for choosing a good inflated structure. This could be very problematic as a given structure can have infinitely many inflations.*

5A) While extremely relevant, this is a far-from-simple question, and answering it goes beyond the scope of the current paper. In the revised manuscript we have highlighted this important open question in the conclusions, as follows: *"A single causal structure has an unlimited number of potential inflations. Selecting a good inflation from which strong polynomial inequalities can be derived is an interesting challenge. To this end, it would be desirable to understand how particular features of the original causal structure are exposed* *when different nodes in the causal structure are duplicated. By isolating which features are exposed in each inflation, we could conceivably quantify the utility for causal inference of each inflation."* We also note that the hierarchy of inflated structure per [arXiv:1707.06476] partly addressed this concern, in that works provides clear directions within the infinite landscape of potential inflations.

6) *There is an inconstancy between expressions in (42) and (43) because the former indicates pairwise independence and the latter indicates jointly independence. The authors should clarify which one is of interest.*

6A) We are grateful to the referee for pointing out this inconsistency. We have substantial revised and reorganized the text in order to introduce joint independence much earlier in the manuscript, immediately subsequent to introducing the definition of an ancestrally independent pair of nodes. Following Eq. (30) in the revised text we now state that *"Ancestral independence is closed under union… Consequently, pairwise ancestral independence implies joint factorizability."* This is then exploited in the definition of ai-expressible sets in the revised text, see Def. 8 and Eq. (44).

7) *In the definition of ancestral independence on page 13, one node should not be an ancestor of the other either. This is reflected in (29), by the definition of An(.) but it seems that it is not correctly described in the text. Also, from the definition of ancestral independence, it is not clear how the conditional case, mentioned below (29) is defined (e.g., does S need to be a subset of common ancestors, etc.). It seems that in that part the authors just considered the original notion of d-separation.*

7A) The referee is raising two concerns in these comments. Firstly, the referee seems to find the textual description of ancestral independence to be different from the definition in the associated equation, namely Eq. (30) in the revised text. We have ensured that the textual description of ancestral independence also emphasizes the definition in terms of d-separation by the null set. The alternative definition we provide, namely "lacking a common ancestor", is also consistent, however. Recall that this manuscript uses the term "ancestor" to include the variable itself. This non-colloquial usage of terminology is explicitly called attention to earlier in the manuscript; see footnote 6.

7B) The referee points out that the conditional case was ill-defined. We have completely removed all mention of the conditional case in the revised text, as it is not needed in any of the example or definitions which follow.

8) *As the authors have mentioned in Section IV-A, the conditions obtained in the proposed method are implied by ancestral independences among the observed variables of the causal structure. It seems that this is in fact the main focus of the work, both for the way inflated structure is defined and the way necessary conditions are obtained. It would be good if the authors explain the use of this idea as one of the main tools used in the proposed technique in the beginning of the paper.*

8A) The referee is correct in pointing out that ancestral independence underlies nearly all of the example in the paper. We take care not to mislead the reader into thinking otherwise. For instance, we in summarize the technique in the introduction we write *"One then looks for sets of variables within the inflated causal structure with disjoint ancestries and writes down the factorization of their joint distribution."* We have substantially revised the text in Section IV to emphasize the primacy of ancestral independence (as opposed to leveraging more general d-separation relations). For instance, we write *"With the exception of Appendix A, in the remainder of this article we will limit ourselves to working with [ai-expressible sets] , and leave the investigation of more general expressible sets to future work."* We also introduce a new theorem summarizing the main tool of the proposed technique, namely Theorem 9. That theorem clearly explains that ancestral independence is the sole ingredient in the most elementary formulation of the Inflation Technique.

9) *At the beginning of Section III-C, it would be helpful that the authors clarify what other types of conditions, besides inequalities, could have been considered as necessary conditions for compatibility.*

9A) We have revised the text as follows: The main text now states that *"these conditions can always be expressed as sets of inequalities,"* and a footnote then clarifies that *"Note that we can include equality constraints for causal compatibility within the framework of causal compatibility inequalities alone; it suffices to note that an equality constraint can always be expressed as a pair of inequalities, i.e. satisfying $x=y$ is equivalent to satisfying both $x≤y$ and $x≥y$. The requirement that a distribution must be Markov (or Nested Markov) relative to a DAG is usually formulated as a set of equality constraints."*

10) *In Definition 7, the intuition behind rule 1 is not clear. Also, considering the definition of expressibility as "the joint distribution can be expressed as a function of distributions over injectable sets", it is not clear why only these two rules are enough for checking expressibility. Also, rules 1 and 2 could have been presented as a method to check expressibility not as the definition.*

10A) We have revised the text to reflect that the two rules should not be understood as a definition for expressibility but rather a sufficient condition for expressibility. We have also amended Rule 1 to make the intuition behind it clearer, by pointing out the explicit construction P(ABC)= P(AC)P(BC)/P(C).

11) *In the proposed method, we restrict ourselves to a.i. expressible sets. But it is not clear how much we are losing by doing so. It would be helpful that at least for one example the general independence (as explained in Section V-A) be considered and the authors explain what we are missing.*

11A) Section V-A of the main text refers the reader to Appendix D for a concrete example where the use of general expressible sets unlocks conclusions impossible from considering ai-expressible sets alone. For general expressible sets to be relevant at all, there must exist a pair of variables within the DAG which are d-separated some other non-empty set of observable nodes. This feature is lacking in the Triangle structure, and hence giving examples of general expressibility would require introducing new causal structure examples. In the interest of narrative flow we have elected to limit the concrete example to Appendix D. As an aside, the alternative formulation of Inflation per [arXiv:1707.06476] includes a graphical preprocessing called "unpacking" which negates any loss of generality in then considering exclusively ai-expressible sets.

12) *The sentence "one rather determines all constraints that such a family must satisfy in order to arise from a joint distribution" on page 17 is vague and requires more elaboration.*

12A) We have rephrased and elaborated: *"We have just described how, from a specified joint distribution, one can solve the decision problem of whether or not it is compatible with a given causal structure. The procedure is to focus on a particular family of marginals (on the images* *of injectable sets) of the given joint distribution, then from products of these, obtain the distribution on each of the ai-expressible sets. Finally, one asks simply whether the family of distributions on the ai-expressible sets are consistent in the sense of all being marginals of a single joint distribution. By analogous logic, the following technique allows one to systematically derive causal compatibility inequalities: find the constraints that any family of distributions on the ai-expressible sets must satisfy if these are to be consistent in the sense of all being marginals of a single joint distribution. Next, express each distribution of this family as a product of distributions on the injectable sets, according to Eq. (45), and rewrite the constraints in terms of the family of distributions on the injectable sets. These constraints constitute causal compatibility inequalities for the inflated causal structure. Finally, one can rewrite the constraints in terms of the family of distributions on the images of the injectable sets, using Corollary 6, to obtain causal compatibility inequalities for the original causal structure."*

13) *Lemma 4 is the main basis of this work, hence (even though from reasonings such as the one below expression (6) it seems correct), it is better that a formal proof for this lemma be presented in the paper.*

13A) We have ordered the paper such that the proof preceded the lemma, instead of follow it. Indeed, the entirety of Sec. III A is a careful buildup towards Lemma 4. Note that Lemma 4 follows immediately from Eq. (10), which follows in turn from Eq. (7). While we believe the proof as currently presented is already robust, we have made many small phrasing changes in and preceding Lemma 4 to further tighten the narrative flow.

14) *The explanations in section IV-B are not quite clear and need improvement. It would be very helpful if the example in appendix B is brought in the main text.*

14A) We have made numerous small change to the main text in Sec. IV B, and we have changed the variable naming conventions to be consistent Appendix B. After much deliberation, we felt that merging Appendix B into the main text would unduly interrupt the narrative, possibly putting readers off from encountering the important concepts introduced in Secs. IV C&D. Nevertheless, Appendix B is clearly referenced in the main text, and per the referee's own comments, that appendix provides an explicit example to any reader wishing to explore the algorithm formulation more concretely.

Response to Referee 2

1A) *There is some problem with the latex in the contents section (the numbering is overlapping with the text)*

1B) LaTeX formatting errors have been resolved.

2) *At least in the pdf version I had access I had no bibliography. Because of that I cannot judge if the bibliography is adequate or not.*

2A) The bibliography was provided to the editors shortly after the initial manuscript submission. The revised submission has been checked to ensure that the bibliography is correctly included.

3) *In page 2 it is said "Causal inference problems arise in a wide variety of scientific disciplines, from sussing out biological pathways to enabling machine learning." I have already seen machine learning applied to causal inference but not the other way around. What the authors mean with the statement that causal inference enables machine learning?*

3A) The machine learning algorithms based on Bayesian Nets as developed by Judea Pearl and others uses belief propagation on (artificial) causal structures. Strong artificial intelligence capable of explaining its own decisions requires counterfactual statements: “If I would have done X than Y would have been more likely”, which requires preliminary resolving causal relationships. Machine learning algorithms which attempt to causally model their environment are admittedly in infancy today. See, for example, [arXiv:1801.04016]. The text has been revised to read *“Causal inference has applications in all areas of science that use statistical data and for which causal relations are important. Examples include determining the effectiveness of medical treatments, sussing out biological pathways, making data-based social policy decisions, and possibly even in developing strong machine learning algorithms.”*

4) *In page 3 the authors divide the problem in 2 varieties: latent variables with or without bounded cardinality. What about observable variables with unbounded cardinalities? It is not clear from the discussion in this part of the paper whether the inflation method will also apply/be useful in such cases. Could the author please clarify that?*

4A) The text has been revised to no longer divide the problem in two varieties. In light of [arXiv:1709.00707] we now understand that latent variables can be always be assumed to have bounded cardinality without loss of generality (whenever the observed variables all have a finite number of potential outcomes, which is the domain of this paper.)

5) *The authors say also in page 3 "Later work, by Clauser, Horne, Shimony and Holt (CHSH) showed how to derive inequalities directly from the causal structure [17]." Is this really the case? CHSH already had this notion of causal structure? My impression is that they derive their inequality resorting to the usual local realism arguments (that only very implicitly use the notion of a causal structure).*

5A) We have rewritten the sentence to read “Later work by CHSH derived inequalities without assuming any facts about quantum correlations; this derivation can retrospectively be understood as the first derivation of a constraint arising from the causal structure of the Bell scenario alone.”

6) *In page 4 "Furthermore, the fraction of causal structures that are interesting increases as the total number of nodes increases." I believe that this is the case as well... but, is there any formal proof of this statement? Or is this just some sort of common sense?*

6A) Revised to *“Furthermore,they provided numerical evidence and an intuitive argument in favour of the hypothesis that the fraction of causal structures that are interesting increases as the total number of nodes increases.”*

7) *Example 1, page 9. I believe this result could be easily strengthened and generalized. Consider a scenario with with n observable variables and latent variables that connect at most n-1 of them. Using*

the appropriate cut inflation it should be easy to prove that perfect correlation between the n variables is not possible with such causal structure. I would guess that the same argument would hold for larger cardinality.

7A) The referee's intuition about generalizing the result is correct. However, the incompatibility of n-wise perfect correlation with latent variables connecting at-most n-1 of observed variables --- regardless of small or large observed cardinality --- is a no-go theorem already subsumed in the work of Steudel and Ay. Example 1 is meant to illustrate how their fundamental result can be recovered with the Inflation Technique. In light of their earlier work, we feel that the generalization of Example 1 proposed by the referee would not add substantive value to the manuscript.

8) *In page 14 example 4. In the previous section the authors used the inflation method to test perfect correlations. The inequality (33) could be used also to test imperfect correlations, for example mixing probability (10) with an iid distribution. It would be nice to make this point here and show what is the critical value of this mixing parameter below which the inequality (33) stops to be violated. (One could also compare this with results from other methods).*

8A) We appreciate this suggestion for adding value to the paper. We have added multiple new new paragraphs at the conclusion of Example 4 to address this question. The formulation of the result in the revised text is slightly different from the question posed by the referee but the parameterized distribution per Eq. (35) in the revised text can be related to mixing perfect correlation with an iid distribution by a linear change of variable. To answer the referee specifically, the critical value of the mixing parameter (below which the inequality ceases being violated) is precisely 1/2.

9) *In page 14 example 5. In the case of probabilities or correlators I can clearly see the utility of the inflation method (it kinds of "convexifies" and non-convex problem). However, for entropies (already a linear problem) it seems a bit of useless complication. It would be useful to clarify that at this example. A related question: can the inflation technique give rise to non-Shannon type inequalities (starting from Shannon ones)? The inflation technique seems a bit related to the original proofs of the first non-Shannon type inequalities (extending the space to include new variables and imposing some linear constraints on this extended space). That would be a nice connection.*

9A) We have adopted the referee's suggestion to clarify the utility of the inflation method for entropic inequalities at Example 5. We have added the text: *"Standard algorithms already exist for deriving entropic casual compatibility inequalities given a causal structure. We do not expect the methodology of causal inflation to offer any computation advantage in the task of deriving entropic inequalities. The advantage of the inflation approach is that it provides a narrative for explaining an entropic inequality without reference to unobserved variables. As elaborated in Sec, V D, this allows us to systematically derive GPT-valid entropic inequalities. A further advantage is the potential of the inflation approach to give rise to non-Shannon type inequalities, starting from Shannon type inequalities; see Appendix E for further discussion."*