Change Notes

July 30, 2023

1 Miscellaneous changes

Changed title, as it was (unintentionally) very close to two citations, both of which were by different authors and unrelated to one other. Three "Decision Theoretic Foundations of Causal ..." is too many.

2 Examiner 3

Examiner's comments are behind bullets, my responses are afterwards.

- adding a paragraph or subsection to summarize the major differences and the motivation to propose something new would be really helpful.

Added such a table at the end of Ch 1.

- some terminologies are never defined (e.g. risks confusion) and some are used before they are defined such as identical responses and repeated response functions, seemingly assuming they are standard.

Rewrote the sentence with the phrase "risks confusion" (which was not being used as a technical term). Emphasised that the first discussion of "repeated response functions" was an informal example to motivate the subsequent development.

- If I enter "repeated response functions causal" in Google Scholar, nothing pops up, contradictory to the claim that "of repeated response functions underpins many common models of causal inference". And "repeated response function" is only defined 2 pages later on page 63; that is, "sequences of conditionally independent and identical response functions (CIIR sequences), a precise term for what we refer to above as "repeated response functions"

Clarified that other causal inference frameworks depend on assumptions equivalent to conditionally independent and identical response functions, not that the assumption itself is frequently made in the same terms as used in the thesis.

- It would be great if future work can move onto causal effect estimation and testing the framework in some experiments and empirical studies and comparing inferential results in this framework to those obtained from potential outcomes framework and causal graph approach, and even the existing decision theoretic approach.

I agree that evaluation is an important direction for future work, both for this and other approaches to causal inferencem, and have made this point more prominent in the conclusion.

Regarding comparing results from different frameworks: while it is limited to a particular question, I have added an extended discussion of Theorem 5.1.21 which points out how structural models may be used to motivate key assumptions for this theorem, and how a similar conclusion can be derived from the use of structural models alone (with some slightly different side conditions).

- Pearl claims the potential outcome and causal graph paradigms are equivalent mathematically, so does Robins on some particular causal graph. I assume this new framework, after incorporating the domain expert knowledge through decision modeling, would lead to the same causality conclusion in the cases where the assumptions in other approaches are satisfied

In Chapter 5 I show how to represent potential outcomes and structural interventional models as decision models. By the procedures outlined there, it's always possible to turn a potential outcomes or structural interventional model into a decision model. Though I haven't proven it, it is likely that the resulting decision model will admit the same conclusions as the original models, as they share critical defining characteristics, namely the action of interventions and the "potential outcomes switch" respectively. I've added to the introduction to Chapter 5 to make this point explicit.

In the updated discussion of precedent (Section 5.1), I explain how a similar result to the one presented in this section can be obtained from standard considerations in the structural causal modelling literature. This is not a case of getting the same conclusion by "translating all of the assumptions" discussed above, but is instead two different ways of arriving at the same conclusion.

3 Examiner 2

- General point: I note (with approval) that the candidate has attempted careful discrimination between different types of mathematical objects by the use of distinct fonts and/or capitalisation. However there are many places where this breaks down, and further close attention is required.

I have read over the text again. In addition to the corrections noted below, several errors and examples of inconsistent notation were corrected, and Section 5.3 was substantially rewritten for greater clarity.

Examiner 2 provided a great deal of detailed feedback. I have noted the feedback in detail where I've made a substantive response, and otherwise simply noted where I accepted the examiner's suggestions, many of which were pointing at errors in the text or inconsistencies in the notation.

- 1, 2-3, 5-24, 27-28, 30-33, 35-36, 38-39, 43-49, 51-54, 57-60, 62-66, 70-71, 75

Accepted

- 4 p.11, middle. Meaning of $x \mapsto x1$ etc is obscure. Also, do we need the arrow from T_i to Y_i^X ?

Clarified that Y^X is a function from $X \to Y$ and $x \mapsto x - 1$ is a function of this type and hence a possible value of Y^X. Agreed that the arrow T to Y_i^X is unnecessary (removed)

- 25 p.42, top line. $(\mathcal{S}; F)$ should be $(\mathcal{F}; F)$?

Rephrased to make it clear that the former notation is intended and that $\mathcal{F}$ refers to a different object.

- 26 p.42, §2.5.5. This looks deep, but is obscure. What is meant by "Choose id X"? What do "part 1" and "part 2" refer to?

Rephrased and clarified what it means to choose an option and labeled procedure steps more clearly

- 29 p.51, Definition 3.3.4. The bullets describe a Boolean algebra, not a complete atomless Boolean algebra—that comes below. In any case this description seems inadequate. The last bullet should say that, given any a A there exists a unique element of A, which we write as ¬a, with the stated properties. And do we not also need to require e.g. $\neg(a \vee b) = (\neg a) \wedge (\neg b)$?

Terminology change accepted, Boolean algebra axioms are correct (in fact, there are more than strictly necessary) and $\leq$ is correct also

- 34 p.65, second paragraph. I am not clear of the need for a probability set. What is the intended interpretation of C, α?

Added a sentence to clarify the nature of the set C, and why it is appropriate here.

- 37 p.68, second line. Should $\mathrm{del}_{D\mathbb{N}\setminus A}$ be $\mathrm{Del}_{D\mathbb{N}\setminus A}$? Also, this property seems to be related to the SUTDA condition, which may be worth mentioning.

Made a pass through the whole document to ensure del notation was consistent, as well as swap and $\mathbb{I}$).

Locality is more directly comparable to SUTDA than IO contractibility, I added a sentence pointing out the connection in the appropriate place.

- 40 p.70, Definition 4.3.12 (similar for Definition 4.3.13). Upper limit of summation should be n. Should "all $\alpha \in C$" be "all $A \subset X$"? Is there a difference between H and $\mathbb{H}$?

Accepted change to summation limits. Quantification for all $\alpha \in C$ is intended, and I cannot find $\mathbb{H}$ anywhere in the Thesis, including diagrams and appendices

- 41 p.70, bottom. Have you switched rows and columns?

Yes, I swapped rows and columns (and I've now switched them back)

- 42 p.71, Definition 4.3.13. This ties together, in the same row, e.g. the outcome of the third appearance of $D = 1$ and the outcome of the third appearance, at some entirely different point of the sequence, of $D = 2$. This seems an odd thing to do and deserves some comment.

Commented that this definition only makes sense when the model is exchange commutative, which allows us to rearrange the order of appearance of variable pairs without changing the model

- 50 p.83, last display. Should $Swap_{\rho-1}$ be $Swap_{\rho^{-1}}$? Should right-hand side $\mathbb{P}_C^{Y|D}$?

Accepted superscript on ρ; RHS is as intended, rewrote the text to clarify that the point of the equation is that the tranformed comb is not generally a comb itself

- 55 p.85, Lemma 4.5.13. In the second diagram, X and Y should be X_i and Y_i . The output H should be omitted. (Ditto for diagram in Theorem 4.5.14).

Accepted wire deletion, added a sentence explaining wire labeling which while correct is admittedly confusing.

- 56 p.86, Theorem 4.5.15. What is the meaning of "latent"?

Replaced "latent" with "not a function of observed variables"

- 61 p.92, first display. Should $do_{B_jj}(v_B)$ be $do_{B_j}(v_{B_j})$

Clarified notation in response to 60; subscripts are mostly as intended

- 67 p.96, Theorem 5.1.16. See point 44. It is important to know whether it is intended that $V_{1\mathbb{N}}$ and $V_{1Q((N))}$ are almost surely equal, or merely have the same distribution.

Notation clarified in response to earlier comments (they are almost surely equal)

- 68 p.97, second text paragraph. This is on the assumption that a suitable Q exists, which is not guaranteed.

Added a sufficient condition for such a Q to exist

- 69 p.98, top line. Is this odd phrasing intended?

Maybe the odd phrasing was intended, I'm not sure. I've slightly rewritten it regardless.

- 72 Definition B.1.1. This is billed as being about transpositions, but is in fact about finite permutations.

Renamed the operation to "set swap" and clarified the relation to transpositions

- 73 Lemma B.1.2. Many notational and other infelicities here.

Errors corrected. Monotone convergence argument was unnecessary and removed.

- 74 Typically ρ will not map $[n]$ onto $[n]$. What then is ρ^n

ρ^n is required to be a finite permutation that agrees with ρ on the first n elements. Amended the definition to this, and added an argument that such a permutation exists in general.

4 Examiner 1

- Chapter 1: While this chapter is generally clear and sensible, I did find that some of the criticisms focused on specific constrained versions of particular theories (particularly in Section 1.1.5), rather than engaging with the most sophisticated, interesting versions of the theories.

I'm not completely sure what the examiner has in mind here with "most sophisticated, interesting versions of the theories". In particular, I show in section 1.1.5 that a cyclic interventional model with interventions on an unobserved variable can capture intuitively reasonable behaviour that is difficult to capture with acyclic models. I don't know what exactly the examiner has in mind as the "most sophisticated, interesting version" of structural interventional models, but this is certainly not the simplest class of structural interventional model available. I've added a few sentences explaining why various kinds of generalised interventions don't address the example as well as the cyclic model does. I also explained how more general intervention classes allow for more general models of how consequences depend on actions but at the same time make it harder to specify how consequences depend on data. This is all presuming that the sophisticated models the examiner had in mind were generalised interventions, and I don't know if this is the case.

- Chapter 2: This chapter provides a highly technical overview of probability theory, including a relatively quick introduction to string diagram notation. As someone who was not previously familiar with that notational system, I would have appreciated a few more examples, rather than moving directly to the maximally abstract definition. The concepts were clear, but their application was not always obvious.

I added three examples of string diagram usage, showing substitution in string diagrans, how string diagrams can compactly express the notion that intervention operations "cut incoming arrows to the intervened variable" (due to Jacobs (2019)), and how string diagrams can graphically express the fact that, in a sequence, a given conditional probability does not change.

- Chapter 3: This chapter is similarly focused on technical exposition, though now on decision problems and the basics of different types of decision theory. As with Chapter 2, some of the technical machinery was introduced very quickly. More importantly, the criticisms of counterfactuals seemed to hinge (in places) on the difficulties in interpreting them in specific cases, rather than general concerns about whether they are meaningful in general.

Note that the discussion of counterfactuals here is just a brief explanation of why this thesis focuses on decision problems rather than counterfactual ones, and not a dismissal of the value of modelling counterfactual problems (I have explained this explicitly). I also rewrote the relevant section of this chapter slightly to indicate that the reason we don't aim to model counterfactual problems is that they are much less common than decision problems, and because we don't want to introduce a convention for modelling counterfactuals in problems that are not strictly about counterfactuals.

- Chapter 4: I did wish that the candidate would have explored the relative strengths and weaknesses of EC (exchange commutativity) and L (locality), however. In particular, it seems to me that L is relatively innocuous, at least in many decision settings. There are certainly contexts in which the impact on X (of my decision) can depend on the impact on Y, but I think that we often have good reasons to think that L holds. (Obviously, it can fail if there is significant learning over time, but any good decision theory should allow for that.) In contrast, EC seems much less plausible, as it points towards a very strong type of exchangeability. It would have been helpful to have more discussion about whether there are ever contexts in which EC might reasonably be expected to hold, or whether there are slight weakenings of EC that "break" the key theorems (i.e., is EC only a problem in its most extreme form?)

I added Section 4. containing additional examples and discussion of reasons for why L might often be acceptable and EC appears to be less often acceptable. I also added a forward reference to Section 4. which offers additional reasons to consider EC implausible.

- Chapter 5: The theorems for both causal graphical models and potential outcomes models show that we can (mostly) derive the standard frameworks if CIIR holds for variables that are never the target of an intervention. The candidate seems to take this result as a concern for the dominant frameworks, but it was not clear why. I agree that CIIR is often unreasonable when it is assumed for all variables, but the ones that are never the target of interventions are presumably part of more "stable" (in some sense) causal relations, and so CIIR seems less problematic for them. In particular, CIIR will presumably never hold of id_C and so it seems more plausible than in the fully abstract formulation of Chapter 4.

It was not my intention to indicate that this feature was problematic in and of itself. I add a clarification that, as both frameworks feature CIIR sequences

with inputs that may be unobserved, the criticisms of the CIIR assumption from Chapter 4 don't apply.

- Chapter 5 continued: Another concern about this chapter is that there were missed opportunities to explore natural connections with existing research in these dominant paradigms. For example, the candidate raises concerns about interventions not being well-defined in terms of actions (see also parts of chapter 1), but then never connects these worries with Spirtes & Scheines (2004), "Causal inference of ambiguous manipulations," which explicitly examined this problem from the causal graphical model perspective. Similarly, recent work by Schölkopf, Kun Zhang, Mooij, and others has tried to develop a basis for causality in terms of "stable conditionals."

I have revised and substantially expanded the discussion of Theorem 5.1.21. In particular, I connect this theorem to the work of Schölkopf, Kun Zhang, Mooij, and others by showing how an interpretation of structural models in terms of "independent conditionals" can justify the assumptions of Theorem 5.1.21. Schölkopf, Kun Zhang, Mooij talk about how the principle of independent conditionals can justify certain approaches to causal structure discovery – here I show how it can also justify conclusions about the consequences of actions.

I added a discussion of Spirtes & Scheines (2004) to chapter 1, where a similar issue was raised, and note in this chapter how inference in the form of Theorem 5.1.21 does not require decomposing ambiguous interventions into a combination of unambiguous interventions (which is the approach explained in Spirtes & Scheines (2004)).