Results on Counterfactual Invariance

Jake Fawkes [1] **Robin J. Evans** [1]

Abstract

In this paper we provide a theoretical analysis of counterfactual invariance. We present a variety of existing definitions, study how they relate to each other and what their graphical implications are. We then turn to the current major question surrounding counterfactual invariance, how does it relate to conditional independence? We show that whilst counterfactual invariance implies conditional independence, conditional independence does not give any implications about the degree or likelihood of satisfying counterfactual invariance. Furthermore, we show that for discrete causal models counterfactually invariant functions are often constrained to be functions of particular variables, or even constant.

1. Introduction

Causality has emerged as an important language to reason about generalisation and invariance in machine learning. These approaches to generalisation have involved — but are not limited to — modelling changing environments within causal graphs (Peters et al., 2016), defining invariant predictors in terms of causal graphs (Bühlmann, 2020) and applying causality to predict invariant conditional distributions (Magliacane et al., 2018).

Counterfactual Invariance has arisen as a promising new causal definition of invariance (Veitch et al., 2021). The idea is that we should aim for predictors which lead to the same outcome had we, contrary to fact, intervened on a spurious part of our data. For example, if we change the name of an actor in a movie review but keep everything else the same, the sentiment of the review should not change. Therefore we want the sentiment analysis model to be invariant to counterfactual perturbations of spurious features.

The goal of this paper is to analyse a variety of definitions of counterfactual invariance and examine what implications they have in causal models. We look at the definition introduced in Veitch et al. (2021) and an alternative based upon distributional equality from Quinzan et al. (2022), considering how they relate to each other and what they imply in terms of graphical structure.

We then turn to the central question of the relationship between invariance and conditional independence. First, we give a set of independences implied by counterfactual invariance that encompass many existing results in the literature. We then turn to the reverse implication, asking if these independences imply that a given variable or function satisfies counterfactual invariance. To this our answer is negative, and we show that independences do not bound the degree of counterfactual invariance, and that counterfactual invariance remains in some sense 'unlikely,' even if the independences hold. This raises questions about what passing conditional independence tests achieves in terms of counterfactual invariance, as well as how best to discuss the benefit conditional independence brings.

Finally we prove that for almost all discrete causal models the only counterfactually invariant functions are functions whose inputs do not causally depend on the intervening variable. This implies that if all variables causally depend on the intervening variable then almost always the only counterfactually invariant functions are constant. This has implications for counterfactual fairness, as in this context many have argued that almost all variables will depend on sensitive attributes (Kusner et al., 2017). These results suggest that in these cases the only counterfactually fair functions will be constant.

2. Background

Notations Throughout we will use Y to denote the variable we are assessing counterfactual invariance in, Z to be the variable we are intervening on, and X to be the remaining covariates. We will use $\mathcal{Y}$, $\mathcal{Z}$ and $\mathcal{X}$ to denote the domains of the variables.

In reference to a causal graph we will use $\mathrm{Pa}(Y)$ to denote the causal parents of Y, $\mathrm{De}(Z)$ for the causal descendants of Z, i.e. those variables, including Z, that can be reached from Z via a directed path, and $\mathrm{Nd}(Z)$ for the variables which are not descendants of Z.

[*]Equal contribution [1]Department of Statistics, University of Oxford. Correspondence to: Jake Fawkes <jake.fawkes@st-hughs.ox.ac.uk>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2023 by the author(s).

Structural Causal Models (SCM) (Peters et al., 2017) A structural causal model is defined as a pair $\mathbb{C} := (\mathbf{S}, P_{\mathbf{N}})$ where $\mathbf{S}$ is a collection of d structural assignments:

$$X_j := f_j(Pa(X_j), N_j), \quad j = 1, ..., d,$$

where N_j is the noise variable for generating X_j and $P_{\mathbf{N}} = P_{N_1,...,N_d}$ is a distribution over the noise variables. In this we will take all the noise variables to be independent[1] and will use $\mathcal{N}$ to refer to their domain. We define $Y(z)$ to be the variable Y when we intervene in the structural causal models to set $Z = z$. Given the noise distribution, we may define the distribution over arbitrary events, even counterfactual ones, as just the sum over noise variables which permit such an event. For example:

$$P(Y(z) = y, Y = y') = \sum_{\mathbf{n} \in \mathcal{S}} P_{\mathbf{N}}(\mathbf{n})$$

where $\mathcal{S} = \{\mathbf{n} \in \mathcal{N} : Y(z, \mathbf{n}) = y \ \& \ Y(\mathbf{n}) = y'\}$.

2.1. Definitions of Counterfactual Invariance

We now present a variety of different definitions of counterfactual invariance, inspired existing definitions in the literature. The first is in terms of almost sure equality and is based upon Veitch et al. (2021):

Definition 2.1. A variable Y satisfies **almost sure counterfactual invariance** (*a.s.-CI*) with respect to Z if:

$$Y(z) \overset{a.s.}{=} Y(z') \text{ for all } z, z'.$$

Counterfactual invariance can also be defined in terms of distributional equality, similar to Quinzan et al. (2022):

Definition 2.2. We say a variable Y satisfies **distributional counterfactually invariance** ($\mathcal{D}$-CI) conditional on some set of variables W, with respect to Z, if:

$$P(Y(z) = y | W = w, Z = z) = P(Y(z') = y | W = w, Z = z),$$

for all z, z' and for almost all y, w.

Our definition deviates slightly from Quinzan et al. (2022) as we enforce conditioning on the intervening variable. We do this to ensure the definition is truly counterfactual in the sense that we are always asking "*what would have happened had we intervened to set $Z = z'$, given that in reality $Z = z$.*" Finally we can define counterfactually invariant functions:

Definition 2.3 (Veitch et al. 2021)**.** A function $f : \mathcal{X} \to \mathcal{Y}$ is **counterfactually invariant** ($\mathcal{F}$-*CI*) if the variable $\hat{Y} := f(X)$ satisfies almost sure counterfactual invariance. That is:

$$f(X(z)) \overset{a.s.}{=} f(X(z')).$$

[1] In fact most results in this paper do not require full independence and can be adapted to less restrictive counterfactual models; see, for example, Robins & Richardson (2010).

Relationships Between Definitions We now give some basic relationships between the definitions, with proofs in the appendix:

Lemma 2.4. *We have that a.s.-CI implies $\mathcal{D}$-CI conditional on any set of variables, but $\mathcal{D}$-CI implies a.s.-CI only if the conditioning set contains the outcome, Y.*

This lemma allows us to equivalently define functional counterfactual invariance in terms of distributional counterfactual invariance, conditional on the function inputs. This is because conditioning on the inputs implicitly conditions on the value of the function.

Corollary 2.5. *A function f is $\mathcal{F}$-CI if and only if $\hat{Y} = f(X)$ is $\mathcal{D}$-CI conditional X.*

This also shows that the definition presented in Kusner et al. (2017) is equivalent to that presented in Veitch et al. (2021).

3. Counterfactual Invariance and the Causal Graph

We first look at how counterfactual invariance relates to the causal graph with the following lemma showing there can be graphical implications of counterfactual invariance:

Lemma 3.1. *If Y is a.s.-CI with respect to Z then there is no edge from Z to Y.*

However, as the following example shows, counterfactual invariance does not lead to the absence of any other edges:

Example 3.1. Suppose we have the following causal model:

$$\begin{aligned} Z &= N_Z \\ X &= 2\mathbb{1}\{Z = 1\} + N_X \\ Y &= \mathbb{1}\{\text{mod}_2(X) = 0\} + N_Y, \end{aligned}$$

where all noise variables are Bernoulli with parameter $\frac{1}{2}$. Under this model Y is almost surely counterfactually invariant but the causal graph is $Z \to X \to Y$. Therefore counterfactual invariance does not imply the absence of any edge besides $Z \to Y$.

Moreover, counterfactual independence can be implied by graph structure alone. Namely when there is no direct causal path from the intervening variable to the outcome:

Lemma 3.2 (Kusner et al. 2017)**.** *If Y is not a causal descendant of Z it is counterfactually invariant.*

Therefore, in some cases, counterfactual invariance can be read directly from the graph.

4. Counterfactual Invariance and Conditional Independence

We now turn to the relationship between counterfactual invariance and conditional independence. This has received

a large amount of attention as it is one of the only testable implications of counterfactual invariance. In this section we first give a selection of conditional independences implied by the different forms of invariance before turning to the reverse implication, can conditional independence tell us anything about the degree of counterfactual invariance? Finally for discrete models we give a more exact classification of the counterfactually invariant functions in terms of $Nd(Z)$.

4.1. Invariance Implies Independence

We first give a selection of independences implied by counterfactual invariance. In order to do we first need the definition of a valid adjustment set:

Definition 4.1. A set of variables S is known as a valid adjustment set for the pair (Z, Y) if:

$$P(Y(z)) = \int P(Y \mid s, z)P(s)ds.$$

This leads us to the following result for distributional counterfactual invariance. The proof expands on that given in Fawkes et al. (2022) and can be found in the appendix.

Lemma 4.2. *Suppose Y is distributionally counterfactual invariant conditional on some set W, such that there is some valid adjustment set $S \subset W$ for (Z, Y). Then we have $Y \perp Z \mid S$.*

We can now apply the relationships between the definitions that were given in Section 2 to get a set of independences implied by the other forms of invariance.

Corollary 4.3. *The following holds;*

- *if Y is a.s.-CI then $Y \perp Z \mid S$ for any valid adjustment set S for (Z, Y);*
- *if f is $\mathcal{F}$-CI then $f(X) \perp Z \mid S$ for any valid adjustment set S for $(Z, f(X))$*

For DAG models Shpitser et al. (2012) showed that valid adjustment sets for (Z, Y) can be classified as sets S which satisfy:

- S blocks all non-directed paths from Z to Y.
- S does not contain any descendants of any nodes on a direct path from Z to Y.

If we apply this graphical characterisation alongside Corollary A.2 we can re-derive all the conditional independence implications given in Veitch et al. (2021). Further, for more general graphs the conditional independences implied by counterfactual invariance can now be read directly from the graph by finding the relevant adjustment sets. Finally, our proof does not require the any measurability results of additional variables as in Veitch et al. (2021), so we may apply these results without having to rely on additional assumptions like discrete Z.

4.2. But Independence Does Not Imply Invariance

The observational implications presented in the previous section have lead to arguments that when we would like a counterfactually invariant predictor, we should regularise for the conditional independences implied by it. Whilst it is true that variables not satisfying these independences cannot be counterfactually invariant, it is not clear if satisfying conditional independences can tell us anything about the degree of counterfactual invariance.

In this section we study the question of how independence affects invariance. We begin with with an example that shows the difficulty of achieving counterfactual invariance through independence alone.

Example 4.1. Suppose the only variables are Z, Y which are binary and follow the DAG $Z \to Y$. Now if we fix the noise variable N_Y to a given value n_Y we have that Y is a deterministic function of Z as:

$$Y = f_Y(Z, n_Y).$$

So as the noise, N_Y, varies it corresponds to f switching between different deterministic functions from $\{0, 1\}$ to $\{0, 1\}$. Because of the finite domains it is possible to enumerate all functions, $f : \{0, 1\} \to \{0, 1\}$, as:

$$f_0(z) = 0 \qquad f_1(z) = 1$$
$$f_2(z) = z \qquad f_3(z) = 1 - z.$$

This allows us to represent any possible structural causal model as distribution over the functions $\{f_0, f_1, f_2, f_3\}$ or more simply as a distribution, $P(\tilde{N}_Y)$, over $\{0, 1, 2, 3\}$ where when $\tilde{N}_Y = i$ we have that:

$$Y = \tilde{f}_Y(Z, i) = f_i(Z) \tag{1}$$

for the new structural function $\tilde{f}_Y$. This is known as the function response framework (Balke & Pearl, 1994). Details of this construction for more general discrete causal models can be found in Section 3.4 of Peters et al. (2017).

Given an observational distribution, $P(Y, Z)$, over Y, Z we can find all possible SCMs that could give rise to this observed distribution simply by finding all possible distributions of $\tilde{N}_Y$ that give $P(Y \mid Z)$ under (1). To do this let $p_{ij} = P(Y = i \mid Z = j)$. Now the set of possible distributions which comply with $P(Y, Z)$ can be written as probability vectors in a single free parameter λ as:

$$\mathbf{a}(\lambda) = \begin{pmatrix} 0 \\ 1 - p_{00} - p_{01} \\ p_{00} \\ p_{01} \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

where $\mathbf{a}(\lambda) = P(\tilde{N}_Y = i)$ and λ is constrained so that $\lambda_{\min} = \max\{0, p_{00} + p_{01} - 1\}$ and $\lambda_{\max} = \min\{p_{00}, p_{01}\}$.

We may now assess the degree of counterfactual invariance across the causal models with the following:

$$P(Y(0) = Y(1)) = \mathbf{a}(\lambda)_0 + \mathbf{a}(\lambda)_1 = 1 - p_{00} - p_{01} + 2\lambda,$$

where we have almost sure counterfactual invariance precisely when this quantity is equal to 1. We can see this is only possible when $p_{00} = p_{01}$, recovering the independence requirement from Section 4.1. Letting $p_{00} = p_{01} = p$ we can see the set of observationally equivalent SCM is now given by:

$$\mathbf{a}(\lambda) = \begin{pmatrix} 0 \\ 1-2p \\ p \\ p \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

where almost sure counterfactual invariance occurs precisely when $\lambda = p$. However, in this case λ can vary between $\lambda_{\min} = \max\{0, 2p-1\}$ and $\lambda_{\max} = p$, giving a whole range of observationally equivalent SCMs which are not counterfactually invariant. Furthermore we can see that there is only one possible counterfactually invariant SCM amongst the infinite observationally equivalent models.

This example demonstrates the difficulty of validating counterfactual invariance from observational data alone, even in the most simple of cases. Further, as we can embed versions of this example within more general structural causal models this leads to more general results:

Proposition 4.4. *For almost sure counterfactual invariance we have that:*

- *No set of independences imply any bound on $P(Y(z) = Y(z'))$, even if Y is binary.*
- *For discrete models, the counterfactually invariant structural causal models have Lebesgue measure zero within the models satisfying the set of independences in Corollary A.2.*

The Lebesgue measure zero result is similar to existing results on the faithfulness assumption in causal inference. The faithfulness assumption is common in causal discovery as it states that any observed independences arise only from graphical structure. A common justification for making this assumption is that for discrete and Gaussian models it is violated on a set of Lebesgue measure zero (Meek, 1995). Our result is similar to this, however in our case it advises against concluding counterfactual invariance from conditional independence alone.

Due to the relationship between definitions this also gives us results for distributional and functional counterfactual invariance. We present the result for functional counterfactual invariance with the distributional result in Appendix A.3.3:

Proposition 4.5. *For non constant functions f we have that:*

- *No independences imply f is counterfactually invariant.*
- *For $f : \mathcal{X} \to \{0, 1\}$ no independences imply any bound on $P(f(X(z)) = f(X(z'))$.*

This suggests that whilst counterfactually invariant functions must satisfy certain conditional independences, these independences give little assurances as to the degree of counterfactual invariance.

Moreover for discrete models we can more precisely characterize the set of counterfactually invariant functions:

Proposition 4.6. *For a given causal model let $\mathcal{F}^{CI}$ be the set of counterfactually invariant functions $f : \mathcal{X} \to \mathcal{Y}$. Supposing all the variables are discrete we have that for almost all structural causal models:*

$$\mathcal{F}^{CI} = \{f : f \text{ is a function of } \mathrm{Nd}(Z) \text{ only.}\}$$

Alternatively stated, the set of causal models that admit a counterfactually invariant function which takes a descendant of Z as an input, has Lebesgue measure zero.

This suggests that when trying to build counterfactually invariant predictors from observational data alone it is almost always impossible to do better than Lemma 3.2, the original and most basic way to ensure counterfactual invariance provided by Kusner et al. (2017).

A further implication of this statement is that if every variable is a descendant of Z then for almost all causal models the only counterfactually invariant functions are constant. In the language of Veitch et al. (2021), this means the variable $X^{\perp}_{Z}$ corresponding to the part of X not causally affected by Z is almost surely constant. Whilst these results are for discrete models it raises questions about which general causal models admit non-constant counterfactually invariant functions, and if these are in some cases 'special exceptions' as opposed to something that happens in general.

5. Conclusion

In this paper we have examined numerous different existing versions of counterfactual invariance. From here our contributions were to analyse the relationship between counterfactual invariance and the causal graph, provide a variety of results on the implications between conditional independence and counterfactual invariance. Finally we showcased that for discrete models counterfactually invariant functions are almost always functions of a select few covariates. This creates difficulties for finding counterfactually invariant functions if many covariates causally depend on the intervening variable.

References

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pp. 46–54. Elsevier, 1994.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Bühlmann, P. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Drton, M., Sturmfels, B., and Sullivant, S. *Lectures on algebraic statistics*, volume 39. Springer Science & Business Media, 2008.

Fawkes, J., Evans, R., and Sejdinovic, D. Selection, ignorability and challenges with causal fairness. In *Conference on Causal Learning and Reasoning*, pp. 275–289. PMLR, 2022.

Gresele, L., Von Kügelgen, J., Kübler, J., Kirschbaum, E., Schölkopf, B., and Janzing, D. Causal inference through the structural causal marginal problem. In *International Conference on Machine Learning*, pp. 7793–7824. PMLR, 2022.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Lojasiewicz, S. Triangulation of semi-analytic sets. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 18(4):449–474, 1964.

Magliacane, S., Van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

Meek, C. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 411–418, 1995.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Quinzan, F., Casolo, C., Muandet, K., Kilbertus, N., and Luo, Y. Learning counterfactually invariant predictors. *arXiv preprint arXiv:2207.09768*, 2022.

Robins, J. M. and Richardson, T. S. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 84:103–158, 2010.

Shpitser, I., VanderWeele, T., and Robins, J. M. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*, 2012.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

A. Proofs

A.1. Definitions of Counterfactual Invariance

Lemma 2.4. *We have that a.s.-CI implies $\mathcal{D}$-CI conditional on any set of variables, but $\mathcal{D}$-CI implies a.s.-CI only if the conditioning set contains the outcome, Y.*

Proof. Firstly an almost sure equality of variables implies they are equal conditional on any event they must be almost surely equal conditional on almost all w for any set W. Hence as almost sure equality implies any form of distributional equality.

Secondly we need to show that distributional counterfactual invariance does not imply almost sure counterfactual invariance, unless W contains the outcome. First suppose W does not contain the outcome, then we can generate Y and Z from the following model:

$$Z \sim Ber\left(\frac{1}{2}\right), \quad U_Y \sim Ber\left(\frac{1}{2}\right), \quad Y := Z \oplus U_Y,$$

where $Ber(p)$ represents a Bernoulli random variable with probability of success p, and $\oplus$ represents addition modulo 2; all other variables are generated independently. Then we have, conditional on any set W:

$$P(Y(z) \mid W = w, Z = z') = P(Y(z) \mid Z = z') = \frac{1}{2}.$$

So that $P(Y(z) \mid W = w, Z = z') = P(Y(z') \mid W = w, Z = z')$ and so we have distributional counterfactual invariance. However we have $P(Y(0) = Y(1)) = 0$, so we do not satisfy almost sure counterfactual invariance.

For the final statement we require that if the conditioning set W contains the outcome variable Y then both forms of counterfactual invariance are equivalent. For this we note that distributional equality implies:

$$P(Y(z) = y \mid W = w, Z = z) = P(Y(z') = y \mid W = w, Z = z),$$

but

$$\begin{aligned} P(Y(z) = y \mid W = w, Z = z) &= P(Y = y \mid W = w, Z = z) \\ &= P(Y = y \mid Y = y', Z = z) \\ &= \mathbb{1}\{y = y'\}. \end{aligned}$$

Therefore for almost all w, z we have that:

$$P(\hat{Y}(z') = y' \mid W = w, Z = z) = 1,$$

implying that $P(Y(z) = Y(z') = 1$ hence $\hat{Y}(z) \overset{a.s.}{=} \hat{Y}(z')$. □

Corollary 2.5. *A function f is $\mathcal{F}$-CI if and only if $\hat{Y} = f(X)$ is $\mathcal{D}$-CI conditional X.*

Proof. We already have that almost sure counterfactually invariance implies distributional counterfactual invariance, so all that remains to show is the other direction. For this we can apply the previous lemma, as if we condition on the inputs of f we also condition on $\hat{Y}$. □

A.2. Relationship between Counterfactual Invariance and Edges in a Causal Graph

Lemma 3.1. *If Y is almost surely counterfactually invariant with respect to Z then there is no edge from Z to Y.*

Proof. We use the definition of Bongers et al. (2021) which says that 'Z is not a parent of Y' is equivalent to there being a measurable function $\tilde{f}_Y$ such that:

$$f_Y(\mathbf{x}, \mathbf{e}) = \tilde{f}_Y(\mathbf{x}_{\setminus Z}, \mathbf{e})$$

for almost all $\mathbf{e}, \mathbf{x}$, where $f_Y(\mathbf{x}, \mathbf{e})$ is the original structural equation for Y and $\mathbf{x}_{\setminus Z}$ is the covariates $\mathbf{x}$ without Z. But this is equivalent to Y being distributionally counterfactual invariant conditional on $\mathbf{e}, \mathbf{x}_{\setminus Z}$ as we can set:

$$\tilde{f}_Y(\mathbf{x}_{\setminus Z}, \mathbf{e}) = \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus Z, z', \mathbf{e}] \tag{2}$$

For some fixed z'.

First we can note that $f_Y(\mathbf{x}_{\setminus Z}, z', \mathbf{e}) = \tilde{f}_Y(\mathbf{x}_{\setminus Z}, \mathbf{e})$ as we have specified all parts of the structural equation model for Y. From here we have:

$$\begin{aligned}
\mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus Z, z', \mathbf{e}] &= \mathbb{E}[Y(z') \mid \mathrm{Pa}(Y) \setminus Z, z', \mathbf{e}] && (3)\\
&= \mathbb{E}[Y(z') \mid \mathrm{Pa}(Y) \setminus Z, z, \mathbf{e}] && (4)\\
&= \mathbb{E}[Y(z) \mid \mathrm{Pa}(Y) \setminus Z, z, \mathbf{e}] && (5)\\
&= \mathbb{E}[Y \mid \mathrm{Pa}(Y) \setminus Z, z, \mathbf{e}] && (6)\\
& && (7)
\end{aligned}$$

Where the third line follows as $Y(z') \perp Z \mid \mathrm{Pa}(Y), \mathbf{e}$ and then we apply distributional counterfactual invariance. □

A.3. Relationship between Counterfactual Invariance and Conditional Independence

A.3.1. Invariance Implies Independence

Lemma A.1. *4.2 Suppose Y is distributionally counterfactual invariant conditional on some set W such that there is some $S \subset W$ that is a valid adjustment set for (Z, Y). Then we have $Y \perp Z \mid S$.*

Proof. As S is a valid adjustment set we have that $Y(z') \perp Z \mid S$. Now we have:

$$\begin{aligned}
P(Y = y \mid S = s, Z = z') &= P(Y(z') = y \mid S = s, Z = z')\\
&= P(Y(z') = y \mid S = s, Z = z)\\
&= \mathbb{E}_{P(W \mid S=s, Z=z)}[P(Y(z') = y \mid W, Z = z)\\
&= \mathbb{E}_{P(W \mid S=s, Z=z)}[P(Y(z) = y \mid W, Z = z)]\\
&= P(Y(z) = y \mid S = s, Z = z)\\
&= P(Y = y \mid S = s, Z = z)
\end{aligned}$$

Therefore as $P(Y = y \mid S = s, Z = z') = P(Y = y \mid S = s, Z = z)$ we have $Y \perp Z \mid S$. □

Corollary A.2. *The following holds;*

- *If Y is a.s.-CI then $Y \perp Z \mid S$ for any valid adjustment set S for (Z, Y)*
- *If f is $\mathcal{F}$-CI then $f(X) \perp Z \mid S$ for any valid adjustment set S for $(Z, f(X))$*

Proof. These hold as any distribution satisfying almost sure counterfactual invariance will satisfy distributional counterfactual invariance conditional on any set W, therefore the result in Lemma 4.2 can be applied for any valid adjustment set S. □

A.3.2. But Independence Does Not Imply Invariance

Proposition 4.4. *For almost sure counterfactual invariance we have that:*

- *No set of independences imply any bound on $P(Y(z) = Y(z'))$, even if Y is binary.*
- *For discrete models, the counterfactually invariant structural causal models have Lebesgue measure zero within the models satisfying the set of independences in Corollary A.2.*

Proof. For any set of variables V we can simply let Z, Y be as in example 4.1 and generate all other variables independently. So we have:

$$Y = \tilde{f}_Y(Z, \tilde{N}_Y) \tag{8}$$

Where if $\tilde{N}_Y$ is equal to i, $Y = \tilde{f}_Y(Z, i) = f_i(Z)$. The distribution $P(\tilde{N}_Y)$ over $\{0, 1, 2, 3\}$ has probability vector which lies in:

$$\mathbf{a}(\lambda) = \begin{pmatrix} 0 \\ 1-2p \\ p \\ p \end{pmatrix} + \lambda \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \tag{9}$$

Where $P(Y(z) = Y(z')) = \mathbf{a}(\lambda)_0 + \mathbf{a}(\lambda)_1$. If we let $p = \frac{1}{2}$ we can see that λ can vary between $[0, \frac{1}{2}]$ with $P(Y(z) = Y(z'))$ taking all values in $[0, 1]$. Therefore we have constructed an example which satisfies all possible independences however $P(Y(z) = Y(z'))$ is unconstrained.

For the second statement we use use a generalised version of the function response framework as is found in Peters et al. (2017) or Gresele et al. (2022) so that the observable distribution over all variables V arises as:

$$P(V) = \sum_{\mathbf{e}} P((\mathbf{n})) \prod_{i=1}^{n} \mathbb{1}\{v_i = f_{i,n_i}(\mathbf{pa}_i)\}, \tag{10}$$

where n_i indexes the possible functions from $\mathrm{Pa}(V_i)$ to $\mathcal{V}_i$. Now we can relate any probability distribution $P((\mathbf{N}))$ with some vector $p \in \Delta^{m-1}$ in the probability simplex where $m = |\mathcal{N}|$ and the entries of the probability vector are $P(\mathbf{N} = \mathbf{n})$.

Now all the conditional independence statements will correspond to polynomial constraints on the vector p as in Drton et al. (2008). That is they will correspond to a set of functions $\{g_i\}_{i=1}^k$ where $g_i : \Delta^{m-1} \to \mathbb{R}$ and k is the total number of constraints where we require:

$$g_i(p) = 0 \tag{11}$$

Now let $\mathcal{M}$ denote the submodel of Δ^{m-1} that lies in the algebraic variety generated by $\{g_i\}_{i=1}^k$, that is the submodel that satisfies the conditional independence relationship. Now take a rational parameterisation of $\mathcal{M}$, so some $G : \Theta^d \to \mathcal{M}$ where Θ^d is some open set in $\mathbb{R}^d$. From previously almost sure counterfactual invariance corresponds to:

$$P(Y(z) = Y(z')) = 1 \tag{12}$$

is a linear function in $p \in \mathcal{M}$ as:

$$P(Y(z) = Y(z')) = \sum_{\substack{\mathbf{n} \\ Y(z,\mathbf{n}) = Y(z',\mathbf{n})}} P(\mathbf{N} = \mathbf{n}). \tag{13}$$

And is so a rational function in $\theta \in \Theta^d$. However from the previous example we know there exists a point $\theta_0 \in \Theta^d$ such that $G(\theta_0)$ is not counterfactually invariant. Therefore as this rational function is non-zero at some value it must be non zero for almost all $\theta \in \Theta^d$ (Lojasiewicz, 1964). This follows as any polynomial is either zero everywhere or on a set of measure zero.

Hence amongst the set of models satisfying the conditional independence set (given by Θ^d) almost all SCM's are not counterfactually invariant. □

Proposition 4.6. *For a given causal model let $\mathcal{F}^{CI}$ be the set of counterfactually invariant functions $f : \mathcal{X} \to \mathcal{Y}$. Supposing all the variables are discrete we have that for almost all structural causal models:*

$$\mathcal{F}^{CI} = \{f : f \text{ is a function of } \mathrm{Nd}(Z) \text{ only}\}. \tag{14}$$

Alternatively stated, the set of causal models that admit a counterfactually invariant function which takes a descendant of Z as an input, has Lebesgue measure zero.

Proof. As for previous proos for a given causal graph $\mathcal{G}$ all structural causal models can be associated with a point in the probability simplex $p \in \Delta^{m-1}$. Now we work with the Lebesgue measure over Δ^{m-1}.

Now suppose $f : \mathcal{X} \to \mathcal{Y}$ is some function which depends on a covariate that is a descendant of Z. We will show that for almost all causal models this function is not counterfactually invariant. To do so we note that the counterfactual invariance of f will again correspond to some polynomial in our probability distribution p. Hence as this polynomial is either zero everywhere or almost nowhere we simply have to find a point p under which f is not counterfactually invariant.

To do this note f takes as input some variable D which is a descendant of Z. Hence there must be two values d, d' such that for a fixed value $\mathbf{x}_{\setminus D}$ of the other covariates $X \setminus D$ the function f changes value. Now we choose p such that V is a deterministic function of Z that flips between d, d' and all other covariates are set to $\mathbf{x} \setminus D$. Therefore as we intervene on Z we change the value of f so under this structural causal model f is not counterfactually invariant, hence f is not counterfactually invariant under almost all structural causal models.

Now we have that for each function f which depends on some descendants of Z it is counterfactually invariant on a set of measure zero. We can take the union of these sets over all functions to get the set of structural causal models for which there exists one counterfactually invariant function which depends on a descendant of Z. As this is the union of finitely many measure zero subsets it is also measure zero we have that almost everywhere the set of counterfactually invariant functions is:

$$\mathcal{F}^{CI} = \{f : f \text{ is a function of } \mathrm{Nd}(Z) \text{ only.}\} \tag{15}$$

□

A.3.3. Distributional Invariance

In the previous section we demonstrated that almost sure independence is not implied or even bounded by any set of independences. However distributional counterfactual invariance can be implied by conditional independence:

Lemma A.3. *If W is a valid adjustment set then $Y \perp Z \mid W$ is equivalent to Y being distributionally counterfactual invariant conditional on W.*

Proof. As W is a valid adjustment set we have that $Y(z) \perp Z \mid W$. Therefore we may write:

$$P(Y(z) \mid W = w, Z = z') = P(Y(z) \mid W = w, Z = z) \tag{16}$$
$$= P(Y \mid W = w, Z = z) \tag{17}$$

For any z, z'. Therefore distributional counterfactual invariance simply corresponds to:

$$P(Y \mid W = w, Z = z') = P(Y \mid W = w, Z = z) \tag{18}$$

Which is exactly the conditional independence. □

However this is a unique subset of distributional counterfactual invariance where the counterfactual distribution is identifiable from the observational distribution. In general this is not possible and so we can recover similar results to the previous case:

Proposition A.4. *Suppose there is some vertex $w \in W$ such that $w \in \mathrm{Pa}(Y) \cap \mathrm{De}(Z)$ and $Y \not\perp w$. The for distributional counterfactual invariance conditional on W we have all the results from almost sure invariance. That is:*

- *No additional independences imply distributional counterfactual invariance conditional on W.*
- *If Y is binary no additional independences are sufficient to bound:*

$$|P(Y(z)=1|W=w, Z=z) - P(Y(z')=1 \mid W=w, Z=z)|.$$

- *For discrete models, the distributional counterfactual invariant structural causal models have Lebesgue measure zero within the models satisfying any set of conditional independences.*

Proof. The first follows as Y is allowed to depend arbitrarily on w we may simply set $Y = w$. Distributional counterfactual invariance conditional on W then becomes almost sure counterfactual invariance of w by lemma 2.4. From here we can apply proposition 4.4 to show the first two results, noting that the only dependence implied so far is between Y and w. This leads to the final result as distributional counterfactual invariance is again a selection of polynomials in our parameter vector. □