# Counterfactual explainability of black-box prediction models

**Zijun Gao** ZIJUNGAO@MARSHALL.USC.EDU
*Department of Data Science and Operations, University of Southern California*
**Qingyuan Zhao** QZ280@CAM.AC.UK
*Statistical Laboratory, University of Cambridge*

**Editors:** Biwei Huang and Mathias Drton

It is crucial to be able to explain black-box prediction models to use them effectively and safely. Most existing tools for model explanations are associational rather than causal, and we use two paradoxical examples in the full version to show that such explanations are generally inadequate.

Motivated by the concept of genetic heritability in twin studies, we propose a new notion called counterfactual explainability for a black-box prediction model $Y$ that maps the inputs $W_1, \ldots, W_K$ to a real value. Let $\mathcal{S} \subseteq [K]$ be a subset of indices for the model inputs. To begin with, when the model inputs $W_1, \ldots, W_K$ are "causally independent" in the sense that they are probabilistically independent and have no causal effects on each other, we propose to define the total counterfactual explainability of $W_{\mathcal{S}}$ to $Y$ as

$$\xi_Y(\vee_{k \in \mathcal{S}} W_k) := \frac{\mathsf{Var}\,(Y(W) - Y(W'_{\mathcal{S}}, W_{-\mathcal{S}}))}{2\mathsf{Var}(Y(W))}, \tag{1}$$

where $W'_{\mathcal{S}}$ is an identically distributed copy of $W_{\mathcal{S}}$ that is independent of $W$. To quantify the strength of interaction, we propose to use the inclusion-exclusion principle. As an example, the explainability of the interaction between $W_1$ and $W_2$ to $Y$ is defined as $\xi_Y(W_1 \wedge W_2) = \xi_Y(W_1) + \xi_Y(W_2) - \xi_Y(W_1 \vee W_2)$. We prove that this can be alternatively expressed using the variance of an interaction contrast as

$$\xi_Y(W_1 \wedge W_2) = \frac{\mathsf{Var}(Y(W'_1, W'_2) - Y(W'_1, W_2) - Y(W_1, W'_2) + Y(W_1, W_2))}{4\mathsf{Var}(Y(W_1, W_2)}, \tag{2}$$

where $(W_1, W_2)$ and $(W'_1, W'_2)$ are independent and identically distributed. More generally, eq. (2) can be extended to higher-order interaction by using the anchored decomposition of a multivariate function. In fact, we show that (1) defines a *unique probability measure* on what we call the "explanation algebra" generated by $W_1, \ldots, W_K$.

When the input factors are causally dependent and obey a causal directed acyclic graph (DAG), our definition of counterfactual explainability can be extended almost effortlessly by replacing an input factor $W_k$ by its basic potential outcomes in (1). For example, in the structural equation model with additive noise, $W_k$ is replaced by its noise variable.

Our counterfactual explainability has three key advantages: (1) it leverages counterfactual outcomes and extends methods for global sensitivity analysis (such as functional analysis of variance and Sobol's indices) to a causal setting; (2) it is defined not only for the totality of a set of input factors but also for their interactions and all objects in the "explanation algebra"; (3) it also applies to dependent input factors whose causal relationship can be modeled by a directed acyclic graph, thus incorporating causal mechanisms into the explanation.

**Keywords:** Causality, Directed acyclic graph, Explainable AI, Global sensitivity analysis.

The link to the full version: https://arxiv.org/abs/2411.01625.