Causal Dependence Plots

Joshua R. Loftus
Department of Statistics
London School of Economics
London, England, UK
j.r.loftus@lse.ac.uk

Lucius E. J. Bynum
Center for Data Science
New York University
New York, NY, USA
lucius@nyu.edu

Sakina Hansen
Department of Statistics
London School of Economics
London, England, UK
s.a.hansen1@lse.ac.uk

Abstract

To use artificial intelligence and machine learning models wisely we must understand how they interact with the world, including how they depend causally on data inputs. In this work we develop Causal Dependence Plots (CDPs) to visualize how a model's predicted outcome depends on changes in a given predictor *along with consequent causal changes in other predictor variables*. Crucially, this differs from standard methods based on independence or holding other predictors constant, such as regression coefficients or Partial Dependence Plots (PDPs). Our explanatory framework generalizes PDPs, including them as a special case, as well as a variety of other interpretive plots that show, for example, the total, direct, and indirect effects of causal mediation. We demonstrate with simulations and real data experiments how CDPs can be combined in a modular way with methods for causal learning or sensitivity analysis. Since people often think causally about input-output dependence, CDPs can be powerful tools in the xAI or interpretable machine learning toolkit and contribute to applications like scientific machine learning and algorithmic fairness.

1 Introduction

This paper develops Causal Dependence Plots (CDPs) to visualize relationships between input variables and a predicted outcome. Motivated by explaining or interpreting AI or machine learning models [8, 16, 17, 36], for simplicity we consider supervised learning, i.e. regression or classification. We also focus on the model-agnostic or "black-box" setting, where the interpreter can query the model but not access its internal structure. Model-agnostic interpretation methods are functionally limited to observing how the model responds to variation in the inputs. While this initial application forms our practical motivation, we emphasize that CDPs are more general.

Simple explanations that focus on one input variable at a time can be powerful tools for human understanding. However, just as with the interpretation of linear regression model coefficients, these simple relationships can be misleading. When varying one input variable, *we must make some choice about what values to use for the other inputs*. CDPs make this choice using an explicit causal model, and to our knowledge this is the first work that does so. We compare CDPs to other state-of-the-art, non-causal explanation methods like the Partial Dependence Plot (PDP) [12], Individual Conditional Expectation (ICE) [14], Accumulated Local Effect (ALE) [3], and Shapley Additive Explanation (SHAP) feature plot [33]. Explanation methods may respect existing causal dependencies between predictors or break them.

Problem statement. If there are causal relationships between predictors but our visualization, interpretation, or explanation method does not respect them the resulting model explanation may be irrelevant or misleading [37, 54]. Such explanations could lead to incorrect decisions for regulating or aligning algorithmic systems, sub-optimal allocations of resources based on model predictions, a breakdown between human feedback and reinforcement learning systems, or other forms of error and

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3579

harm. In scientific machine learning—where explanations can be used to generate hypotheses for follow-up investigation—a flawed interpretation may support spurious hypotheses. For these reasons, *the causal validity of model explanations should be a top priority*.

High level proposal. We wish to interpret or explain a given supervised machine learning model $\hat{f}(\mathbf{x})$ with p input features $\mathbf{x} = (x_1, \ldots, x_p)$. Specifically, we want to understand how the predictions $\hat{y} = \hat{f}(\mathbf{x})$ of this model depend on feature x_j for a given j, $1 \leq j \leq p$. PDPs and ICE plots do this by varying x_j and holding the other features $\mathbf{x}_{\setminus j}$ constant, where $\mathbf{x}_{\setminus j}$ is the $(p-1)$-tuple containing all features except for x_j. This implicitly assumes independence between x_j and the other inputs. Our method replaces this independence assumption with an Explanatory Causal Model (ECM) for the input features. This auxiliary ECM is a tool we use to help explain $\hat{f}$, it determines how other inputs $\mathbf{x}_{\setminus j}$ vary when x_j is changed.

CDP pseudo-algorithm. To construct a CDP showing how $\hat{f}(\mathbf{x})$ depends on x_j, a user specifies an ECM $\mathcal{M}$ containing the predictors $\mathbf{x}$, and an intervention $I(x_j)$ in $\mathcal{M}$. The intervention changes x_j, and may change other features if they are caused by x_j in $\mathcal{M}$. The type of intervention is chosen based on the type of causal explanation desired, with several example options demonstrated later. An explanatory dataset $\mathcal{D} = \{\mathbf{x}_i : i = 1, \ldots, n\}$ can be given or, if unavailable, generated by $\mathcal{M}$. Note that we use the notational convention where i indexes observations or examples while j indexes features. The horizontal axis of the plot is specified by a grid $\{\tilde{x}_{j,k} : k = 1, \ldots, K\}$ of possible values for x_j, with k indexing grid points. For each observation $\mathbf{x}_i$ in $\mathcal{D}$, and at each grid point $\tilde{x}_{j,k}$:

1. Use the ECM to simulate counterfactual values $\mathbf{x}^*_{i,k}$ for all features of observation i under the intervention $I(\tilde{x}_{j,k})$.
2. Input counterfactual features to the prediction function $\hat{f}$, and store the resulting counterfactual prediction $\hat{y}^*_{i,k} = \hat{f}(\mathbf{x}^*_{i,k})$ in an array indexed by (i, k).

For each observation i in $\mathcal{D}$, construct the individual counterfactual prediction curve $(\tilde{x}_{j,k}, \hat{y}^*_{i,k})$ by connecting points that are adjacent on the plot grid. Plot the empirical average of these curves, which is the main output of the CDP. The individual curves can be shown or suppressed as desired. *The resulting CDP shows how the model's predictions $\hat{y}$ causally depend on x_j when this predictor is varied by the intervention $I(x_j)$ in ECM $\mathcal{M}$.*

PDP and ICE algorithm. Start with the notation and setup as above but without any ECM or intervention. For each observation $\mathbf{x}_i$ in $\mathcal{D}$, and at each grid point $\tilde{x}_{j,k}$:

1. Define $\mathbf{x}'_{i,k}$ by setting feature x_{ij} to the grid point $\tilde{x}_{j,k}$ and keeping other features $\mathbf{x}_{\setminus j}$ fixed at their original values in $\mathbf{x}_i$ from $\mathcal{D}$, that is
$$\mathbf{x}'_{i,k} := (x_{i1}, \ldots, x_{ij} \leftarrow \tilde{x}_{j,k}, \ldots, x_{ip}). \tag{1}$$
2. Compute prediction $\hat{y}'_{i,k} = \hat{f}(\mathbf{x}'_{i,k})$ and store in an array indexed by (i, k).

Plotting $(\tilde{x}_{j,k}, \hat{y}'_{i,k})$ generates an ICE curve for each i, and the empirical average of these is the PDP.

Motivating example. Consider a model for parental income P, school funding F, and graduates' average starting salary S, with ECM shown in the bottom row of Figure 1. In the top row, the ECM functions are plotted in the left panel, and the remaining panels show visual explanations of supervised models that predict $\hat{S} = \hat{f}(P, F)$. In this example the training data for black-box models was generated by the ECM, but later we will see real data examples where this is not the case. Blue curves show how $\hat{S}$ depends on P when P is causally manipulated *without* holding F constant, i.e. under the intervention do$(P = p)$. Orange curves show the dependence of $\hat{S}$ on P when F is held constant at its observed value, and coincide exactly with standard PDPs. Full definitions of these are given in Section 2. Several key takeaways:

- Comparing the direct (or partial) dependence curves and total dependence curves we see *there can be qualitative differences depending on the type of explanation desired*, for example one can be increasing while the other is decreasing. This is a consequential fact when considering how interventions may change (predicted) outcomes. Increasing P causes larger values of $\hat{S}$, but if the increase in P is done while holding F constant then it could cause a decrease in $\hat{S}$ (or a smaller increase). **An intervention which does *not* hold other**

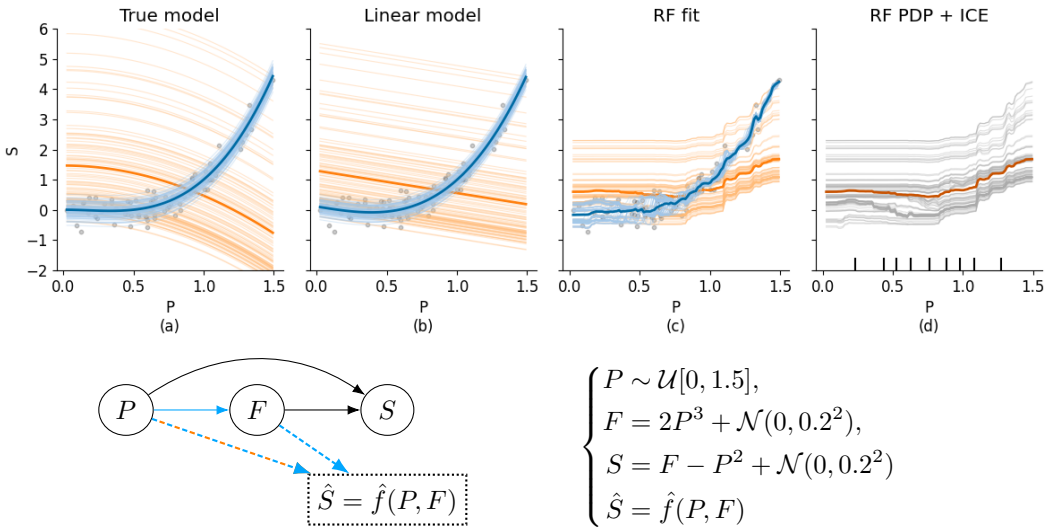

Figure 1: Motivating example. Causal Dependence Plots (top row) and the Explanatory Causal Model (bottom) for the motivating example. Points show the explanatory dataset, which in this example is also the training data for the predictive models. Counterfactual curves for individual points are shown as thin, light lines, with averages displayed as thick, dark lines. Total Dependence (TDP) is represented in blue and Natural Direct Dependence in orange. Panel (a) shows the relationships of the ECM. Panels (b-c) show CDPs for a linear model and random forest (RF) model, respectively. Panel (d) shows PDP and ICE curves for the RF model from a standard software library. This is identical to our NDDP in panel (c). We show this holds true in general: PDP/ICE are a special case of CDPs.

predictors constant—arguably the canonical causal operation—can be shown by our TDP.

- *Our framework includes some existing model explanation plots like ICE and PDPs as special cases*. In panels (c-d), and later in Theorem 2.10, we see that **PDP** + **ICE** = **NDDP**. A practitioner seeing only the PDP in panel (d) may conclude that "dependence" of $\hat{S}$ on P is weak, especially if $P \leq 1$. The TDP in panel (c) shows a stronger increasing relationship closer to the true total dependence and a more holistic view of how $\hat{S}$ depends on P. Our work clarifies that the weaker form of dependence shown by PDPs is natural direct dependence. We also see the same weak dependence empirically for SHAP and ALE plots in Figure 4.
- *Explanations of models can be qualitatively different from the underlying causal relationships*. For example, the random forest in panel (c) shows a direct dependence of $\hat{S}$ on P that is increasing when the true direct dependence of S on P is decreasing. Predictive machine learning models may fail to capture causal dependence, and in this case studying the black-box would not necessarily help us learn about the real world. As another example, panel (b) shows that the total dependence of a linear model on a predictor can be non-linear, in this case because the mediator F depends non-linearly on P.

1.1 Applications

Different combinations of the predictive setting and choice of ECM generate many uses for CDPs. In general, ECMs can be designed based on a particular desired explanation, make use of prior domain knowledge, or be learned and estimated from data using causal learning methods in a modular fashion. Importantly, the ECM does not need to contain the outcome variable y except in one special case—residual plots—to be discussed later.

Causal bridge. In one special case we may wish to understand how $\hat{f}$ depends on a variable z *which is not one of* $\hat{f}$*'s input features* but is causally related to them. Other methods cannot do this, but CDPs can provided the ECM also contains z. For simplicity we choose notation in the rest of the paper to reflect the case where the explanatory variable is an input, but this is not a loss of generality

https://doi.org/10.52202/079017-3579

since we can simply define $\hat{f}_{\text{drop}}(\mathbf{x}, z) := \hat{f}(\mathbf{x})$ and apply CDPs to $\hat{f}_{\text{drop}}$. Since the ECM may vary $\mathbf{x}$ when z is changed, we can see how $\hat{f}$ depends on z. *This could be useful to probe a predictive model for fairness with respect to a sensitive attribute that the model does not use directly*.

Incomplete causal knowledge. There are various applications where an ECM does not need to be a fully specified or "correct" model for all features. First, CDPs only use the predicted—and not actual—outcome. This is useful for semi-supervised or anticausal learning: given causal structure *among predictors only* and a supervised learning model, attempt causal inference for the outcome [52, 63]. Second, we may only require explanations or plots for one or a small number of features. In such cases *we only need information from the ECM about interventions on the features of interest and their causal descendants*, and not other predictors. Finally, predictive models often use features that are known transformations or representations of inputs, and these transformations can be used to construct an ECM. For example, if the features are (x, x^2, z), an ECM can simply encode the fact that x^2 depends causally on x. Even if we do not know the dependency between x and z, we can make use of our partial knowledge about the features.

Multiparty auditing, e.g. for fairness. An owner of a predictive model may not have causal knowledge or incentives to use such knowledge. Predictive accuracy is their only concern. But a separate party, like a regulator, may audit that model. This party may have more causal knowledge due to specializing in auditing, or may be legally obliged to make certain causal assumptions for the purpose of the audit, e.g. to allow disparities only along certain causal pathways and not others. Previous work applied causality to fairness [5, 10, 27–29, 32, 34, 38, 48, 60, 62], recourse [7, 26, 44, 57], and other desiderata. *Existing methods like PDPs are limited to only showing direct dependence, and this may hide the full extent of unfairness or discrimination* [18]. CDPs can be used to probe a black-box for unfairness in the form of total dependence or partially controlled dependence.

Explanations under covariate shift. Often a pre-trained model is used for predictions on data from a different data generating process than the training DGP. We can use ECMs and CDPs to visualize how the model will behave out-of-distribution. ECMs could even be chosen adversarially.

Scientific theory development. Large and complex models may be fit to data where underlying structure is largely unknown. In such settings, relatively simple ECMs can be used to formulate simple hypotheses relating some predictors and plot causal dependencies to check these hypotheses or generate new ones. This can also be done hypothetically, assuming an ECM for exploration.

1.2 Contributions

After defining the CDP framework we demonstrate CDPs on synthetic and real datasets in Section 3 and Appendix B, including in conjunction with structural causal learning in B.2. We compare CDPs with other state of the art competitor visualization methods, for example in Figure 4. Theorem 2.10 establishes the first universally valid causal interpretation of PDP and ICE plots. Finally, in Section 2.7 we illustrate how to visualize uncertainty about the choice of ECM.

2 Methodology

2.1 Supervised learning models

We are given a predictive model $\hat{f}$, possibly estimated or learned using empirical risk minimization (ERM) $\hat{f} = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell\left(h(\mathbf{x}_i), y_i\right)$ with some loss function ℓ, pre-specified function class $\mathcal{H}$, and an independent and identically distributed training sample $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ with feature vectors $\mathbf{x}_i^T \in \mathbb{R}^p$. In Section 2.6 we focus on simple mediation analysis and partition the predictor variables into subsets so that X and M both notate predictors, M being a mediator.

2.2 Fundamental problem of univariate explanations

To create an explanation of model dependence on a single feature, like a plot with x_j on the horizontal axis and $\hat{f}$ on the vertical axis, *we must decide what to do with the other features* when varying x_j along the plot axis. Most explanation methods use the same approach as the PDP and ICE plots: they *hold other features fixed* at values in a (auxiliary, explanatory) dataset. This may be unrealistic

if other features depend on x_j causally, or even mathematically undefined if features are mutually constitutive, e.g. (x, x^2) or the set $\{\text{GDP}, \text{GDP per capita}, \text{population}\}$.

2.3 Structural Causal Models

Our notational conventions and definitions are influenced by [6, 41, 43]. Let $\mathbf{U}$ be a set of exogeneous noise variables, $\mathbf{V}$ a set of $p = |\mathbf{V}|$ observable variables, and $\mathbf{G}$ a set of functions such that for each $j \in 1, \ldots, p$ we have $V_j = g_j(\mathbf{PA}_j, U_j)$, where $\mathbf{PA}_j \subseteq \mathbf{V}$ and $U_j \subseteq \mathbf{U}$ are the observable and exogeneous parents, respectively, of variable V_j. Let the directed acyclic graph (DAG) $\mathcal{G}$ have vertices given by variables and, for each $V_j \in \mathbf{V}$ and each of the parent variables in $\mathbf{PA}_j$ and U_j, a directed edge oriented from the parents to V_j.

Definition 2.1 (Structural Causal Model (SCM))**.** A (probabilistic) SCM $\mathcal{M}$ is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{G}, P_{\mathbf{U}} \rangle$ where $P_{\mathbf{U}}$ is the joint distribution of the exogeneous variables. This distribution and the functions $\mathbf{G}$ determine the joint distribution $P^{\mathcal{M}}$ over all the variables in $\mathcal{M}$. Finally, causality in this model is represented by additional assumptions that $\mathcal{M}$ admits the modeling of interventions and/or counterfactuals as defined below.

Definition 2.2 (Interventions)**.** For the SCM $\mathcal{M}$, an intervention I produces a modified SCM denoted $\mathcal{M}^{\text{do}(I)}$ which may have different structural equations $\mathbf{G}^I$. Correspondingly, some variables may have different parent sets, so the DAG representation $\mathcal{G}^{\text{do}(I)}$ may also change. We denote the new, interventional distribution as $P^{\mathcal{M};\text{do}(I)}$. A simple class of interventions involves intervening on one variable, e.g.

$$I = \text{do}\left(V_j \coloneqq \tilde{g}(\tilde{\mathbf{PA}}_j, \tilde{U}_j)\right),$$

which changes how V_j and all variables on directed paths from V_j in $\mathcal{G}$ are generated. An even simpler sub-class of these are the atomic interventions setting one variable V_j to one constant value v, which we denote $I_{j,v} := \text{do}(V_j = v)$. Note that in this case V_j has no parents in the graph $\mathcal{G}^{\text{do}(I)}$; the source of the intervention itself is outside the world of the model.

Interventions are useful for modeling changes to a data generating process (DGP), for example, experiments that control a particular variable to see how its value changes other variables, or policy changes aimed at altering or removing existing causal relationships. In addition to generating new observations as a DGP, an SCM can also be used to model counterfactual values for observations that have already been determined. A counterfactual distribution is an interventional distribution defined over a specific dataset with information or constraints given by some of the observed values in that data, as we now describe.

Definition 2.3 (Counterfactuals)**.** Let $\mathbf{V}$ be the observed variables for observations in a given dataset, $\mathbf{PA}_j = \mathbf{v}$ and U_j the observed and exogeneous parents of variable V_j, and I and intervention that modifies any of V_j's parents. The intervention I may hold some or all of $\mathbf{v}$ fixed and vary $U_j \leftarrow u$, passing these through $g_j(\mathbf{v}, u)$, or through $\tilde{g}_j(\tilde{\mathbf{v}}, u)$ if the intervention also changes any of $\mathbf{v} \leftarrow \tilde{\mathbf{v}}$. The counterfactuals $V_j(\tilde{\mathbf{v}}, u)$ are values V_j would have taken if any of its observed and/or exogeneous parents had taken the different values $(\tilde{\mathbf{v}}, u)$. To define the counterfactual distribution $P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)}$, we use the posterior or conditional (depending on our probability model approach) distribution $P_{\mathbf{U}|\mathbf{V}=\mathbf{v}}$ to model uncertainty about $\mathbf{U}$ while computing counterfactual values of variables for an observation in the modified SCM $\mathcal{M}^{\text{do}(I)}$.

Remark 2.4. Note that if the desired causal explanation uses counterfactuals, then we likely obtain the observed values from an auxiliary explanatory dataset. But since an SCM can generate data, we may also use it to generate the initial observed values and then re-use these when computing counterfactuals for the explanation.

2.4 Causal explanations

Our proposed solution to the fundamental problem highlighted for univariate explanations is to use an auxiliary ECM $\mathcal{M}_j$ and let this causal model determine how other features vary as functions of x_j. We denote these explanations as $\mathcal{E}_j(\hat{f}; \mathcal{M}_j)$ or $\mathcal{E}_j(\hat{f})$ if the context is clear. In the deterministic or noiseless case, suppose we know continuous functions g_{kj} such that $x_k = g_{kj}(x_j)$, with g_{jj} the identity. In this case the model $\mathcal{M}_j$ tells us $g(x_j) := (g_{1j}(x_j), \ldots, x_j, \ldots, g_{pj}(x_j))$ is a curve in $\mathbb{R}^p$ parameterized by x_j, and we generate the explanation $\mathcal{E}_j(\hat{f})$ by plotting $\hat{f}\left(g(x_j)\right)$ against x_j.

 https://doi.org/10.52202/079017-3579

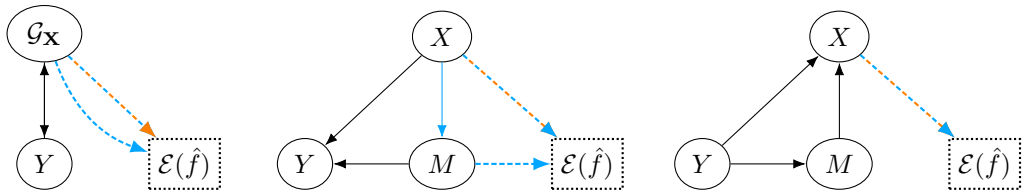

Figure 2: An ECM for predictors is used to produce an explanation $\mathcal{E}(\hat{f})$ of the predictive model $\hat{f}$. Solid arrows represent possible causal relationships in the ECM, and dotted arrows show dependence of the model explanation on predictors. In (a) $\mathcal{G}_{\mathbf{X}}$ denotes the subgraph of the SCM for predictors. In the mediation example (b), predictor X causes Y directly and also through mediator M, creating an important distinction between direct dependence (orange) and total dependence (blue). The reverse causality example (c) shows variables useful in predicting Y may be caused by Y, and also be causes of prediction $\hat{Y} = \hat{f}(X, M)$ and the explanation of that prediction.

To extend this strategy to non-deterministic causal models we use an ECM $\mathcal{M}_{\mathbf{X}}$ for the predictor variables with $\mathcal{G}_{\mathbf{X}}$ its associated DAG. We represent this graphically in Figure 2. The expressive power of SCMs allows us to pose various interpretive questions and compute various kinds of explanations by performing operations in $\mathcal{M}_{\mathbf{X}}$.

2.5 Causal Dependence Plots

For the following definitions, we assume predictor variables $\mathbf{X} \in \Omega_{\mathbf{X}}$, an outcome of interest $Y \in \Omega_Y$, and a black-box function $\hat{f}(x) : \Omega_{\mathbf{X}} \to \Omega_Y$ with outputs that we may also denote $\hat{Y}$. A structural causal model $\mathcal{M}$, either assumed or learned from data, specifies causal relationships for the predictors $\mathbf{X}$, i.e. it need not involve the outcome Y. Note that predictors may only be a subset of the variables in $\mathcal{M}$ as in the "causal bridge" application discussed in Section 1.1.

Definition 2.5 (Explanatory Causal Model (ECM))**.** An ECM $\mathcal{M}'$ augments the SCM containing predictors by including the predicted outcome $\hat{Y}$ as an additional variable with $\hat{f}$ as its structural equation.

Generating causal explanations for $\hat{f}$ involves performing abduction, action, and prediction with this ECM. In a large ECM graph we may suppress all arrows into $\hat{Y}$ except those from the explanatory feature and its descendants. This is to simplify the display, as in Figure 5.

Additional notation and conventions. We use the shorthand $\hat{f}(P^{\mathcal{M}})$, where $\hat{f}$ takes a distribution $P^{\mathcal{M}}$ as its argument, to denote using data from that distribution as the input to the black-box function $\hat{f}$. For each type of causal explanation with a given *Named Effect* based on intervention I, we define the *Individual Counterfactual Named Effect* curves as the set of counterfactual curves $\hat{f}(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)})$ for each individual, the *Named Effect Function* as their (empirical) expectation $\hat{\mathbb{E}}\left[\hat{f}(P^{\mathcal{M}|\mathbf{V}=\mathbf{v};\text{do}(I)})\right]$, and the *Named Dependence Plot* as a plot displaying all of these curves.

Definition 2.6 (Causal Dependence Plot (CDP))**.** Given a function $\hat{f}$, explanatory dataset $\mathcal{D}$, ECM $\mathcal{M}$, and family of interventions I_θ parameterized by θ, we construct a plot with θ as the horizontal axis and display Individual Counterfactual (IC) curves

$$\mathsf{IC}(\theta) = \hat{f}(P^{\mathcal{M}|\mathbf{X}=\mathbf{x};\text{do}(I_\theta)}). \tag{2}$$

These show the effect of intervention I_θ on black-box output for each individual observation in the explanatory dataset as θ varies. The (empirical) average of these (over the explanatory data) is (an estimate of) the Causal Effect Function (CEF), and a plot showing the IC and CEF is a Causal Dependence Plot.

We typically apply this to create plots for one explanatory feature X_s at a time using interventions like $I_\theta = \text{do}(X_s = \theta)$. Horizontal axes for plots use a grid over the possible values of X_s given by its range in dataset $\mathcal{D}$. Bar graphs can be used when the explanatory feature is categorical.

Next is perhaps the most straightforward and important named effect.

Definition 2.7 (Total Dependence Plot (TDP))**.** For an intervention I, the Individual Counterfactual Total Effect (ICTE) curves

$$\mathsf{TE}(I) = \hat{f}(P^{\mathcal{M}|\mathbf{X}=\mathbf{x};\text{do}(I)}) \tag{3}$$

show the total effect of intervention I on black-box output for each individual observation in the explanatory dataset. The (empirical) average of these (over the explanatory data) is (an estimate of) the Total Effect Function (TEF), and a plot showing the ICTE and TEF is a Total Dependence Plot (TDP). We compute the TDP following Algorithm 1.

Algorithm 1 Total Dependence Plot (TDP)

Inputs: $\mathcal{M}$ (ECM), $\hat{f}$ (black-box predictor), $\mathcal{D}$ (explanatory dataset), X_s (covariate of interest)

Let X be a grid of possible values of X_s
Set N to the number of observations in $\mathcal{D}$
Initialize $N \times |X|$ matrix of predictions $\hat{Y}$
for x in X **do**
  Define intervention $I = \text{do}(X_s = x)$
  Sample counterfactual dataset $\mathcal{D}_{X_s \leftarrow x}$ entailed by $P^{\mathcal{M}|D;\text{do}(I)}$
  Set $\hat{Y}[:, x]$ to $\hat{f}(\mathcal{D}_{X_s \leftarrow x})$
end for
Plot N lines $(X, \hat{Y}[i, :])$ {(Individual Counterfactuals)}
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ {(Causal Dependence)}

Remark 2.8. In the remaining definitions, we give notation only for the individual counterfactual curves and leave the other objects implicitly defined.

We often wish to decompose how much of the total effect of X on $\hat{Y}$ is attributable to different pathways between the variables. This can be explored via direct dependence below, as well as with other named CDPs described in Appendix A.

Definition 2.9 (Natural Direct Dependence Plot (NDDP))**.** Given intervention I define a corresponding intervention J that intervenes on all descendants of any variables that are changed by I, except for $\hat{Y}$, and resets them to their observed values in dataset $\mathcal{D}$. We then define the Individual Counterfactual Natural Direct Effect curves

$$\mathsf{NDE}(I) = \hat{f}(P^{\mathcal{M}|\mathbf{X}=\mathbf{x};\text{do}(I,J)}). \tag{4}$$

This quantity represents the effect of intervention I on black-box output $\hat{Y}$ while all variables not directly intervened upon by I are fixed at their 'natural,' i.e., pre-intervention values in $\mathcal{D}$. Algorithm 4 demonstrates how to compute the NDDP.

From this construction of NDDP, we see by comparing it to the **PDP and ICE algorithm** that it is equivalent to these, confirming what we observed in Figure 1(d).

Theorem 2.10 (PDP + ICE = NDDP)**.** *When generating plots for the predictive model $\hat{f}$ using the dataset $\mathcal{D}$ and feature X_s, the ICE plot curves and Individual Counterfactual Natural Direct Dependence curves are identical. Hence, the NDDP is identical to a PDP that includes ICE curves.*

Proof. We have implicitly assumed both plots will use the same range for their horizontal axes. This is natural as implementations use the range of the feature in the dataset, and both plots are constructed from the same feature X_s in the same dataset $\mathcal{D}$. Since the PDP and NDDP both contain empirical averages of their respective Individual curves it suffices to show these are equal at each point $\tilde{x}$ in the plot grid.

Consider individual i in dataset $\mathcal{D}$ with features $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. The value of the ICE curve at $\tilde{x}$ for this individual is, from (1), $\hat{f}(\mathbf{x}'_{i,k})$ where $\mathbf{x}'_{i,k} := (x_{i1}, \ldots, x_{is} \leftarrow \tilde{x}, \ldots, x_{ip})$, i.e. the original features $\mathbf{x}_i$ but with entry s set to $\tilde{x}$. We must show this is the same as the value of the Individual Counterfactual Natural Direct Dependence curve at $\tilde{x}$ for individual i. Applying Definition 2.9, we use interventions

$$I = \text{do}(X_s = \tilde{x}) \text{ and } J = \text{do}(X_j = \mathbf{x}_j \text{ if } X_s \text{ is an ancestor of } X_j).$$

 https://doi.org/10.52202/079017-3579

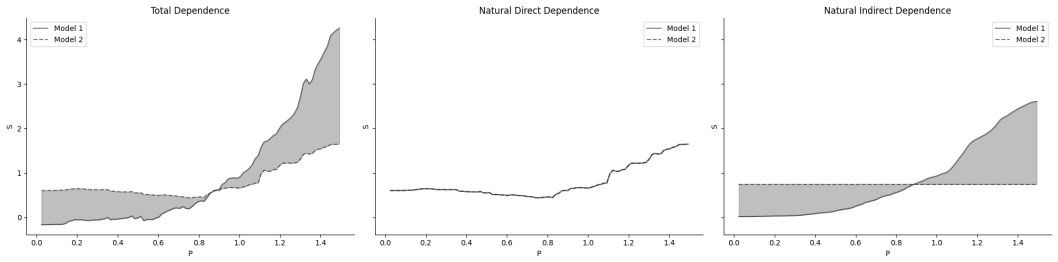

Figure 3: TDP, NDDP, and NIDP uncertainty bands for the salary example using the random forest model in Figure 1. The range of curves is induced by two candidate ECMs described in Section 2.7.

The NDDP applies these in the order I followed by J. First, I sets the value of feature X_s to $\tilde{x}$ for all individuals, and may modify other features if they are descendants of X_s. Then J intervenes on each descendant X_j of X_s and resets it to its observed values in $\mathcal{D}$, and in particular for individual i these are each reset to x_{ij}. Hence, the value of the Individual Counterfactual Natural Direct Dependence curve at $\tilde{x}$ for individual i is also given by $\hat{f}(\mathbf{x}'_{i,k})$. □

Remark 2.11. Note that there is some subtlety in the assumption of using the same dataset: a Bayesian probability modeling approach to SCMs may add more randomness when sampling counterfactuals. In this case, rather than the ICE and ICNDD curves being identical, the ICE will equal the expectation of the ICNDD curves over this additional source of randomness. The additional randomness is specified by priors over exogenous variables, and the expectation can be estimated by more sampling at a computational cost of a constant factor. Since the usage of these plots is visual and somewhat qualitative this subtlety is not an important limitation of the theorem, and it would only apply under particular modeling assumptions.

Remark 2.12. To our knowledge this is the first result establishing a universally valid causal interpretation of PDPs. **Its most important limitation is that it applies to the model output $\hat{Y}$ and not necessarily the original outcome Y.**

Several other types of named CDPs are described in Appendix A.

2.6 Mediation analysis

Many applications involve a causal structure we refer to as a mediation triangle, with examples shown in Figure 1 and Figure 2b. In mediation analysis, we often wish to decompose how much of the total effect of X on Y is attributable to the pathway through M and how much of it is direct. CDPs allow us to visualize frequently studied quantities of interest in this setting including other special cases defined in Appendix A.1. Although mediation analysis motivates CDPs and helps build intuition, we emphasize that our definitions can be used *with any structural causal model*. See Section 3 for other, more complex examples.

2.7 Uncertainty and sensitivity analysis

There are various ways to incorporate uncertainty about the ECM into CDPs. We explore a natural first extension of the CDP that shows a *range between possible effect functions* induced by a *set of auxiliary ECMs*. The set of ECMs could be pre-specified or, for example, given by a Markov equivalence class output by a causal structure learning algorithm. Returning to our motivating example from Section 1, we might question whether parental income P impacts school funding F, considering instead an SCM without mediation: $P \to S \leftarrow F$. Figure 3 shows a range of possible effect functions interpolating between this ECM without the indirect effect and the original ECM in Section 1, for each of the TDP, NDDP, and NIDP. In this we have assumed the same structural equations for the edges that are common to both models. These plots show a range for how $\hat{S}$ may depend on P when we are unsure how F depends on P. Figure 9 in Appendix B.2 shows an example with real data where we use candidate ECMs discovered by the PC algorithm. These examples are not confidence regions, but any method for producing confidence sets in SCMs could also be used with CDPs to display uncertainty regions. Future work can develop additional methods for

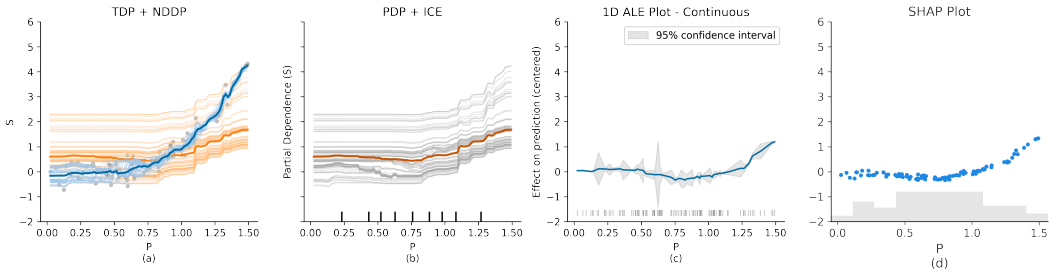

Figure 4: Comparison of CDP (a) with PDP (b), ALE (c), and SHAP plots (d) for the salary example in Figure 1. Our TDP stands out, and all other plots are qualitatively similar.

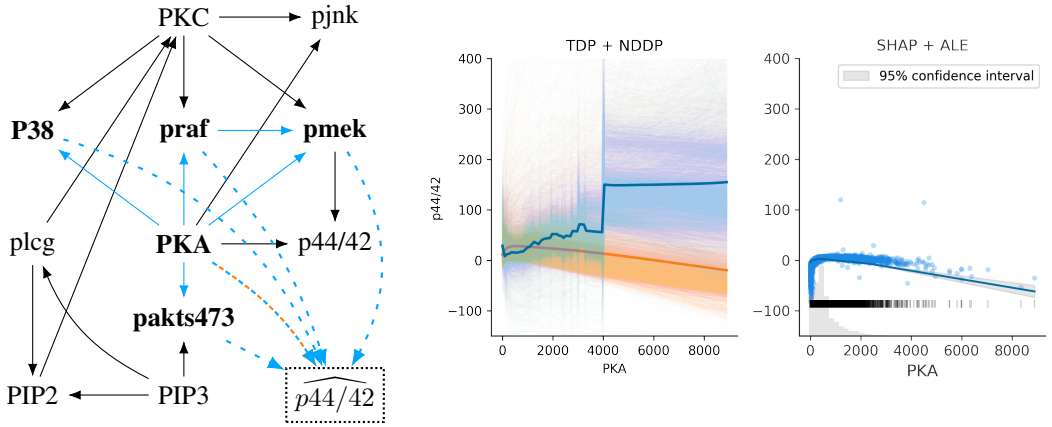

Figure 5: ECM for the Sachs et al. [49] dataset (left), CDPs for an MLP predictive model (center), ALE (line) and SHAP (points) plots (right). All plots visualize the effect of PKA on predicted p44/42. PKA and its descendants are bolded. The NDDP (i.e. PDP + ICE), ALE, and SHAP all show an overall decrease, while the TDP shows an increase. *Conclusions depend strongly, qualitatively, on the specific interpretive question we ask, and causal modeling allows us to formulate questions precisely*.

visualizing uncertainty, for example by leveraging sensitivity analysis based on conformal prediction [9, 21, 31, 61].

3 Experiments

We demonstrate CDPs in a series of experiments with simulated and real datasets.

Comparison with other explanatory plots. Figure 4 shows accumulated local effect (ALE) [3] and Shapley Additive Explanation (SHAP) [33] feature plots for the salary example. These appear similar to the PDP, with a more weakly increasing relationship than that seen in the TDP. TDPs represent a significant and novel contribution to the existing model visualizations.

Real data with domain knowledge. An ECM may be constructed using domain expertise. Figure 5 shows an ECM and CDPs for the Sachs et al. [49] dataset of expression levels of proteins and phospholipds in human cells, for which data and a ground-truth DAG[1] are publicly available in the Causal Discovery Toolbox [24]. While the actual biology of the problem is not our focus here, there are meaningful takeaways from the figure. For this model, the TDP shows an increasing relationship, while the NDDP/PDP shows a decrease. *The overall direction of the trend in predictions based on PKA is reversed if we hold other predictors fixed.* This is an important lesson for using model explanations in scientific machine learning.

[1]Following the discussion in [45] and follow-up ground truth DAG for the Sachs et al. [49] dataset in Figure 5 of [45], we choose the edge PIP3 → PIP2 in order to eliminate a would-be cycle.

https://doi.org/10.52202/079017-3579

Additional experiments can be found in Appendix B. Notably, in Appendix B.2 we *learn an ECM from data with a causal structural learning algorithm* and then use it to produce CDPs. The main takeaway of the real data experiments is that CDPs can be useful in practice.

In simulation experiments we know the true DGP, so we can compare its functional form to various black-box models and their explanatory plots. Results in Appendix B.1 show that CDPs are sensitive to whether the functional form assumptions of the black-box model fit the DGP. In other words, *if a black-box model is a poor fit to the DGP, then CDPs can accurately explain the black-box but will not reflect the true DGP*. This limitation is not specific to CDPs but applies to all explanation methods. Figure 6 also shows that different ECMs can produce different CDPs for the same black-box model, and *a misspecified ECM can produce misleading CDPs*.

4 Discussion

Related work. Recent work in recourse [25] uses contrastive or counterfactual explanations [56]. Some of this focuses on causal dependence [50]. Blöbaum and Shimizu [4] identify the predictor with the largest total effect, which is most applicable when assuming linearity. Zhao and Hastie [63] investigated causal interpretations of PDP, aiming for causal inference about the underlying DGP, and showed that when the DGP satisfies the backdoor criterion [39] then a PDP visualizes the total effect (TE) of a predictor. Cox Jr [11] observed an association between partial dependence plots and NDE, an equivalence we formally establish in Theorem 2.10, to our knowledge the first such result. Lazzari et al. [30] weight observations when computing PDPs. There has been some recent work creating causal variants of SHAP [1, 13, 19, 23, 58], and in future work we will explore comparisons of appropriate special cases of CDPs with these. We are not aware of any previous causal explanation work with the generality of CDPs.

Limitations. Causal modeling always involves some limitations [15, 47]. For CDPs, full specification of an ECM can be a strong assumption. However, in Sections 1.1 and 2.7 we discussed some ways this can be relaxed. In general, *if a causal explanation is desired or necessary, then we cannot avoid making causal assumptions*. Model-agnostic explanation methods also always have certain limitations [2, 36]. For example, if the predictive model fails to fit the DGP, then any model explanation will also fail if our interpretive goal is to learn about the DGP [63]. *CDPs may be misleading if the true DGP differs in important ways from the ECM*, as shown in Figure 6. However, standard PDPs and similar explanation methods also require auxiliary explanatory data, and that data may also differ from the target DGP. So this is not an additional limitation specific to our method.

Conclusion. Causal Dependence Plots use an explanatory causal model to create plots with causal interpretations. This allows us to use the powerful language of structural causal models to pose and answer a variety of meaningful questions. Our framework generalizes Partial Dependence Plots, which Theorem 2.10 shows it includes as a special case, and allows other kinds of causal interpretations we have not seen previously explored in the literature. Future work in this direction could expand on some canonical causal structures for useful applications, or interface with other kinds of models, for example extending to non-tabular data by applying causal representation learning. Relating explanation methods to Pearl's ladder of causation [40], most previous interpretable machine learning and explainable AI methods—like PDPs—concern associations and hence are confined to the first rung of the ladder. With CDPs we ascend the ladder, creating model interpretations intended to change the world. Interpretability provided the initial motivation for CDPs, but since plots are qualitative CDPs also open the door for future work on causal methodology that relaxes assumptions while maintaining *visual validity*.

Broader Impacts. There are many potential societal consequences of our work: essentially those shared by all tools for model interpretability and explainability. Model explanations can be misleading, either due to error or intentional deception. When a user is convinced by a flawed model explanation to reach misguided conclusions about a model, they may make harmful or sub-optimal decisions about how or whether to use that model. For example, if an explanation tool is used to assess the fairness of a model, a flawed explanation could lead to the conclusion that a discriminatory model is fair or that a fair model is discriminatory. In applications related to science, a flawed model explanation can lead to wasting resources pursuing a dead-end hypothesis or to missing out on an important discovery. Similarly, poor business decisions can be made on the basis of flawed explanations.

References

[1] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1054–1070, 2022.

[2] T Altmann, J Bodensteiner, C Dankers, T Dassen, N Fritz, S Gruber, et al. *Limitations of interpretable machine learning methods*, chapter 3-4. 2020. URL `https://slds-lmu.github.io/iml_methods_limitations/`.

[3] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

[4] Patrick Blöbaum and Shohei Shimizu. Estimation of interventional effects of features on prediction. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

[5] Lucius Bynum, Joshua Loftus, and Julia Stoyanovich. Disaggregated Interventions to Reduce Inequality. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13. Association for Computing Machinery, New York, NY, USA, October 2021. ISBN 978-1-4503-8553-4. URL `https://doi.org/10.1145/3465416.3483286`.

[6] Lucius Bynum, Joshua Loftus, and Julia Stoyanovich. Counterfactuals for the future. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[7] Lucius E.J. Bynum, Joshua R. Loftus, and Julia Stoyanovich. A new paradigm for counterfactual reasoning in fairness and recourse. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7092–7100. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/784. URL `https://doi.org/10.24963/ijcai.2024/784`. Main Track.

[8] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, August 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832. URL `https://www.mdpi.com/2079-9292/8/8/832`. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[9] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021.

[10] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[11] Louis Anthony Cox Jr. Modernizing the bradford hill criteria for assessing causal relationships in observational data. *Critical reviews in toxicology*, 48(8):682–712, 2018.

[12] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[13] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in neural information processing systems*, 33:1229–1239, 2020.

[14] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[15] Sander Greenland and Mohammad Ali Mansournia. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European journal of epidemiology*, 30:1101–1110, 2015.

[16] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):93:1–93:42, August 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

[17] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021. ISSN 2689-5595. doi: 10.1002/ail2.61. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61.

[18] Sakina Hansen and Joshua Loftus. Model-agnostic auditing: a lost cause? In *CEUR Workshop Proceedings*, volume 3442, 2023.

[19] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

[20] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55. URL https://matplotlib.org/. Licensed under the PSF license.

[21] Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

[22] D. Jomar, S. Galli, and S. Kiran. PyALE: A python implementation of Accumulated Local Effects. 2024. URL https://github.com/DanaJomar/PyALE. Licensed under the MIT License.

[23] Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, pages 10476–10501. PMLR, 2022.

[24] Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *Journal of Machine Learning Research*, 21(37):1–5, 2020. URL https://github.com/FenTechSolutions/CausalDiscoveryToolbox. Licensed under the MIT License.

[25] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv:2010.04050 [cs, stat]*, March 2021. URL http://arxiv.org/abs/2010.04050. arXiv: 2010.04050.

[26] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Towards Causal Algorithmic Recourse. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Lecture Notes in Computer Science, pages 139–166. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04083-2. doi: 10.1007/978-3-031-04083-2_8. URL https://doi.org/10.1007/978-3-031-04083-2_8.

[27] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

[28] Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making Decisions that Reduce Discriminatory Impacts. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3591–3600. PMLR, May 2019. URL https://proceedings.mlr.press/v97/kusner19a.html. ISSN: 2640-3498.

[29] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[30] Matilde Lazzari, Jose M Alvarez, and Salvatore Ruggieri. Predicting and explaining employee turnover intention. *International Journal of Data Science and Analytics*, 14(3):279–292, 2022.

[31] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B*, 83(5):911–938, 2021.

[32] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal Reasoning for Algorithmic Fairness. *arXiv:1805.05859 [cs]*, May 2018. URL `http://arxiv.org/abs/1805.05859`. arXiv: 1805.05859.

[33] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. URL `https://github.com/shap/shap`. Licensed under the MIT License.

[34] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

[35] Olvi L Mangasarian and William H Wolberg. Cancer diagnosis via linear programming. Technical Report 5, 1990.

[36] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2022. URL `https://christophm.github.io/interpretable-ml-book/`.

[37] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[38] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[39] Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993. ISSN 08834237. URL `http://www.jstor.org/stable/2245965`.

[40] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[41] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL `https://scikit-learn.org/`. Licensed under the BSD License.

[43] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

[44] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350. Association for Computing Machinery, New York, NY, USA, February 2020. ISBN 978-1-4503-7110-0. URL `https://doi.org/10.1145/3375627.3375850`.

[45] Joseph Ramsey and Bryan Andrews. Fask with interventional knowledge recovers edges from the sachs model. *ArXiv*, abs/1805.03108, 2018.

[46] Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=pOoKI3ouv1`.

[47] Paul Rosenbaum. *Observation and experiment: An introduction to causal inference*. Harvard University Press, 2017.

[48] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/1271a7029c9df08643b631b02cf9e116-Abstract.html.

[49] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005. doi: 10.1126/science.1105809. URL https://www.science.org/doi/abs/10.1126/science.1105809.

[50] Numair Sani, Daniel Malinsky, and Ilya Shpitser. Explaining the behavior of black-box prediction algorithms with causal learning. *arXiv preprint arXiv:2006.02482*, 2020.

[51] Moritz Schauer. Causalinference.jl: Causal inference, graphical models and structure learning in julia. 2022. URL https://github.com/mschauer/CausalInference.jl. Licensed under the MIT "Expat" License.

[52] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012.

[53] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020. URL https://github.com/py-why/dowhy. Licensed under the MIT License.

[54] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021.

[55] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[56] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9:11974–12001, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3051315. Conference Name: IEEE Access.

[57] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287566. URL https://doi.org/10.1145/3287560.3287566.

[58] David Watson. Rational shapley values. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1083–1094, 2022.

[59] William Wolberg. Breast Cancer Wisconsin (Original). UCI Machine Learning Repository, 1992. URL https://doi.org/10.24432/C5HP4Z. Licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

[60] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal Intersectionality and Fair Ranking. In Katrina Ligett and Swati Gupta, editors, *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 7:1–7:20, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-187-0. doi: 10.4230/LIPIcs.FORC.2021.7. URL https://drops.dagstuhl.de/opus/volltexte/2021/13875. ISSN: 1868-8969.

[61] Mingzhang Yin, Claudia Shi, Yixin Wang, and David M Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pages 1–14, 2022.

[62] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[63] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021. doi: 10.1080/07350015.2019.1624293. URL https://doi.org/10.1080/07350015.2019.1624293.

 https://doi.org/10.52202/079017-3579

A Methodology supplement

A.1 Additional CDPs

For completeness we include here definitions for several other important, named CDPs.

Definition A.1 (Partially Controlled Dependence Plot (PCDP))**.** Consider intervention I affecting a subset of variables in $\mathcal{G}_{\mathbf{X}}$ and atomic intervention C that holds constant a disjoint subset of variables. The Individual Counterfactual Partially Controlled Effect curves

$$\mathsf{PCE}(I, C) = \hat{f}(P^{\mathcal{M}|\mathbf{X}=\mathbf{x};\text{do}(I,C)}) \tag{5}$$

represent the effect of intervention I on black-box output $\hat{Y}$ while other variables are set (via intervention) to specific constant values. We compute the PCDP via Algorithm 3.

Definition A.2 (Natural Indirect Dependence Plot (NIDP))**.** Consider atomic intervention I and a corresponding intervention K that removes from $\mathcal{G}_{\mathbf{X}}$ all outgoing edges from any of the nodes intervened upon by intervention I and sets those nodes to their observed values in the explanatory dataset. For example, if $I = \text{do}(A = a, B = b)$, then intervention K will remove all outgoing edges from A and B and set A and B to their original observed values. We then define Individual Counterfactual Natural Indirect Effect curves

$$\mathsf{NIE}(I) = \hat{f}(P^{\mathcal{M}_{\mathbf{X}}^{\text{do}(I)}|\mathbf{X}=\mathbf{x};\text{do}(K)}). \tag{6}$$

Notice that intervention I is performed before intervention K. This quantity represents the effect of intervention I on black-box output $\hat{Y}$ that is due only to any indirect pathways to $\hat{Y}$. We compute the NIDP following Algorithm 5. The difference between two values of this function can be used to express the natural indirect effect as a special case.

A.2 Algorithms

This section describes additional algorithms including several special cases of named CDPs.

Algorithm 2 Explanatory Causal Model (ECM)

Inputs: $\mathcal{M}$ (SCM), $\hat{f}$ (black-box predictor), $\mathbf{S} \subseteq \mathbf{X}$ (covariates used by black-box)
Output: $\mathcal{M}'$ (SCM)

Make copy $\mathcal{M}'$ of SCM $\mathcal{M}$ and perform all subsequent operations on this copy
Add node for $\hat{Y}$ to causal graph $\mathcal{G}'$ of SCM $\mathcal{M}'$
for x in $\mathbf{S}$ **do**
  Add edge in $\mathcal{G}'$ from x to $\hat{Y}$
end for
Set structural equation for node $\hat{Y}$ to $\hat{f}$
Set exogenous variable $U_{\hat{Y}}$ to 0

Algorithm 3 Partially Controlled Dependence Plot (PCDP)

Inputs: $\mathcal{M}$ (ECM), $\hat{f}$ (black-box predictor), $\mathcal{D}$ (explanatory dataset), X_s (covariate of interest), C (intervention controlling other variables in $\mathcal{M}$)

Let X be a grid of possible values of X_s
Set N to the number of observations in $\mathcal{D}$
Initialize $N \times |X|$ matrix of predictions $\hat{Y}$
for x in X **do**
  Define intervention $I = \text{do}(X_s = x, C)$
  Sample counterfactual dataset $\mathcal{D}_{s\leftarrow x,C}$ entailed by $P^{\mathcal{M}|D;\text{do}(I)}$
  Set $\hat{Y}[:, x]$ to $\hat{f}(D_{s\leftarrow x,C})$
end for
Plot N lines $(X, \hat{Y}[i, :])$ {(Individual Counterfactuals)}
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ {(Causal Dependence)}

Algorithm 4 Natural Direct Dependence Plot (NDDP)

Inputs: $\mathcal{M}$ (ECM), $\hat{f}$ (black-box predictor), $\mathcal{D}$ (explanatory dataset), X_s (covariate of interest)

Let X be a grid of possible values of X_s
Set N to the number of observations in $\mathcal{D}$
Initialize $N \times |X|$ matrix of predictions $\hat{Y}$
Get all descendants of X_s in $\mathcal{M}$, excluding $\hat{Y}$, and store in $\mathbf{C}$
Get observed values of all variables in $\mathbf{C}$ and store in $\mathbf{c}$
Define intervention $J = \text{do}(\mathbf{C} = \mathbf{c})$
for x in X **do**
  Define intervention $I = \text{do}(X_s = x)$
  Sample counterfactual dataset $\mathcal{D}_{X_s \leftarrow x}$ entailed by $P^{\mathcal{M}|D;\text{do}(I,J)}$
  Set $\hat{Y}[:, x]$ to $\hat{f}(\mathcal{D}_{X_s \leftarrow x})$
end for
Plot N lines $(X, \hat{Y}[i, :])$ {(Individual Counterfactuals)}
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ {(Causal Dependence)}

Algorithm 5 Natural Indirect Dependence Plot (NIDP)

Inputs: $\mathcal{M}$ (ECM), $\hat{f}$ (black-box predictor), $\mathcal{D}$ (explanatory dataset), X_s (covariate of interest)

Let X be a grid of possible values of X_s
Set N to the number of observations in $\mathcal{D}$
Initialize $N \times |X|$ matrix of predictions $\hat{Y}$
Get all descendants of X_s in $\mathcal{M}$, excluding $\hat{Y}$, and store in $\mathbf{C}$
Make copy $\mathcal{M}'$ of SCM $\mathcal{M}$
for x in $\mathbf{C}$ **do**
  Remove all incoming edges to x from $\mathcal{M}'$
end for
Define intervention $I = \text{do}(X_s = x)$
for x in X **do**
  for i in N **do**
    Sample counterfactual observation d_c for unit i entailed by $P^{\mathcal{M}|D[i];\text{do}(I)}$
    Get counterfactual values of all variables in $\mathbf{C}$ from observation d_c and store in $\mathbf{c}_i$
    Define intervention $J = \text{do}(X_s = x, \mathbf{C} = \mathbf{c}_i)$
    Sample counterfactual observation d'_c for unit i entailed by $P^{\mathcal{M}'|D[i];\text{do}(J)}$
    Set $\hat{Y}[i, x] \leftarrow d'_c[y]$ for index y corresponding to node $\hat{Y}$
  end for
end for
Plot N lines $(X, \hat{Y}[i, :])$ {(Individual Counterfactuals)}
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ {(Causal Dependence)}

B Experiments supplement

B.1 Model misspecification

Consider the non-linear mediation example with the following DGP:

$$
\begin{aligned}
U_X, U_{M_1}, U_{M_2}, U_Y &\sim \mathcal{N}(0, 1) \\
X &= U_X \\
M_1 &= \frac{1}{2} X^3 + U_{M_1} \\
M_2 &= \frac{1}{4} X^3 + U_{M_2} \\
Y &= M_1^2 + M_2^2 - \frac{1}{2} X^2 + U_Y .
\end{aligned}
\tag{7}
$$

https://doi.org/10.52202/079017-3579

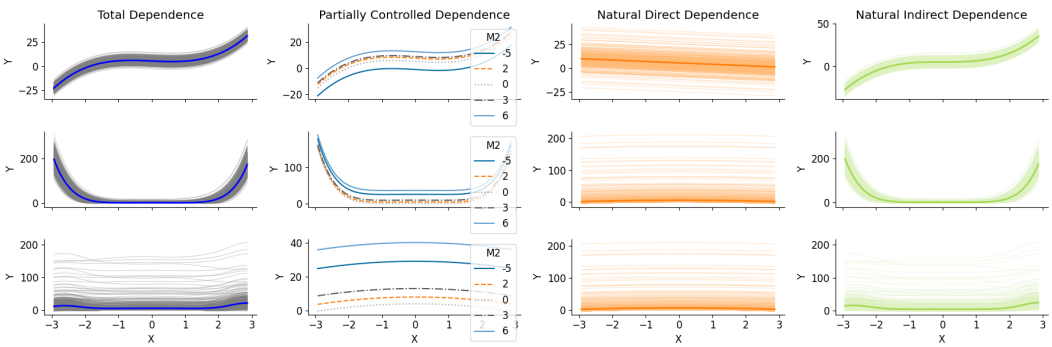

Figure 6: CDPs for the simulation example with data from (7), shown for a 'good' black-box and correct ECM (top row), a 'bad' black-box model and correct ECM (middle row), and a 'good' black-box and misspecified ECM (bottom row).

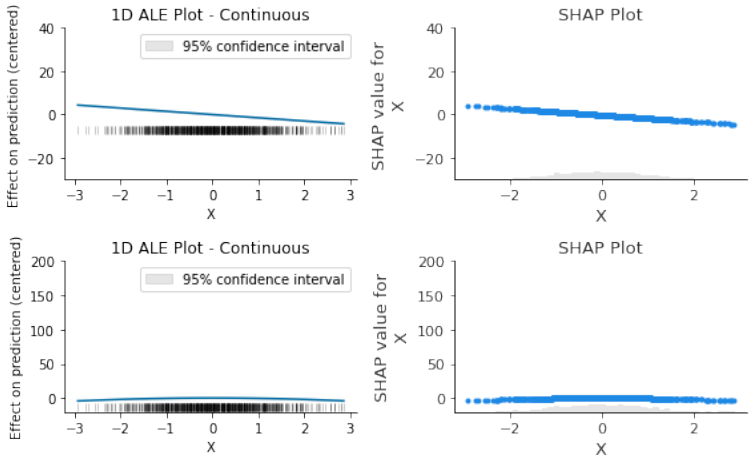

Figure 7: ALE and SHAP for the simulation example with data from (7), shown for a 'bad' black-box model (top row), and a 'good' black-box (bottom row). Both ALE and SHAP show a similar relationship as NDDP for both models.

We use this DGP to fit two different black-box models: one model that assumes the correct functional form (i.e., the relationship for Y shown in the DGP above), and an 'incorrect' model that predicts Y via linear regression. We use two different ECMs to construct CDPs, one which is the true DGP and one which incorrectly assumes the structure $M_1 \rightarrow X \rightarrow M_2$. Figure 6 shows the CDPs for each of these models using the black-box training data as the explanatory data. Similarly, Figure 7 shows SHAP and ALE plots using the same explanatory data.

We can glean a couple insights from Figure 6. First, CDPs are sensitive to whether the functional form assumptions of the black-box model fit the ground truth data generating process. Good explanations for $\hat{Y}$ may be different from good explanations for Y if the black-box model is poorly specified. The second is that a misspecified ECM can produce bad explanations even when the black-box model correctly fits the true causal relationships in the DGP.

In Figure 7 we see that ALE and SHAP produce similar explanations to the PDP/NDDP. We saw this previously for the random forest black-box model in the example from Figure 1.

B.2 Real data with structural causal learning

The Breast Cancer Wisconsin (Original) dataset [35] is a publicly available dataset often used to test algorithms on medical data. The dataset contains 9 ordinal variables, which represent attributes of the cells within a breast mass: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape,

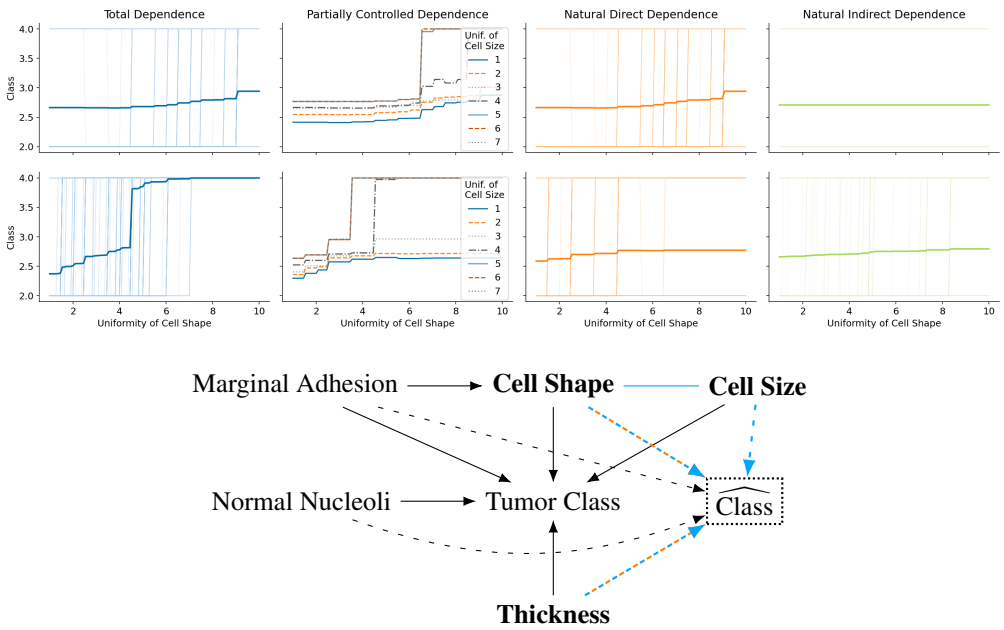

Figure 8: Breast cancer data example. CDPs for a random forest classifier and predictors Clump Thickness (first row) and Uniformity of Cell Shape (second row). Structural graph $\mathcal{G}_B$ for the ECM learned by the PC algorithm (last row). The outcome Class is binary: 2 for benign, 4 for malignant. For the undirected edge between Cell Size and Cell Shape, we investigate the sensitivity to the different options in Figure 9. Note: Our intention is not to make conclusive scientific statements, but only to demonstrate how CDPs could be used in conjunction with causal structure learning.

Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses. The outcome variable is the class of the breast tumor, benign or malignant.

We use a causal structural learning algorithm, specifically the PC algorithm [55] implemented in Julia `CausalInference` [51], to learn a DAG for this dataset, on a smaller subset of predictor variables for simplicity. Figure 8 shows the resulting DAG and CDPs for a random forest model to classify the Class variable.

This shows CDPs can be combined with other causal methods like structural learning algorithms. The PC algorithm output had an undirected edge between Cell Size and Cell Shape. We next explore the graph structures consistent with this uncertain edge.

Figure 9 shows the TDP, NDDP, and NIDP for a learned additive noise model with three different structures consistent with $\mathcal{G}_B$: (1) with the edge Cell Shape $\to$ Cell Size, (2) with the edge Cell Size $\to$ Cell Shape, and (3) with no edge between Cell Size and Cell Shape. This figure shows that the takeaway about cell shape impacting tumor class is indeed sensitive to our choice about the uncertain edge, particularly for the TDP.

B.3 Individual curves showing heterogeneity

One criticism of PDPs is that they may hide heterogeneity or individual variation, and for this reason it may be good practice to include the ICE curves in any PDP. Just like the individual curves in an ICE plot, the individual counterfactual curves in our CDPs can show important effects that are hidden by averaging as illustrated in Figure 10. In our implementation the default settings for CDPs—and our recommendation—is to show these individual curves.

B.4 Causal dependence for residuals

CDPs are applicable not only to understand model outputs but also to understand model performance. For example, CDPs can be used to probe residuals (or other measures of error) under distribution

https://doi.org/10.52202/079017-3579

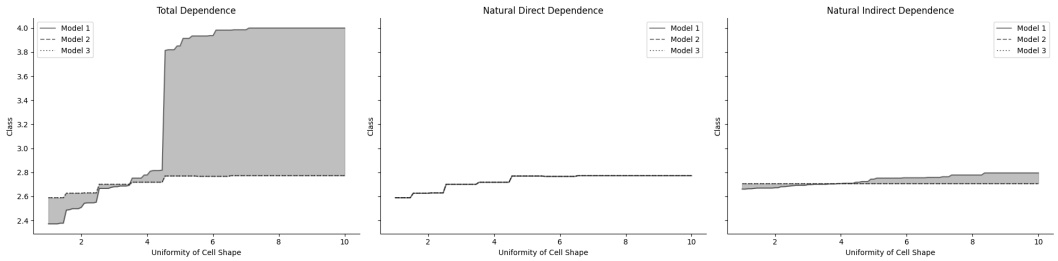

Figure 9: Total Dependence Plots, Natural Direct Dependence Plots and Natural Indirect Dependence Plots for the Breast Cancer Wisconsin dataset under three possible DAGs consistent with the PC algorithm output: (1) with the edge Cell Shape $\rightarrow$ Cell Size, (2) with the edge Cell Size $\rightarrow$ Cell Shape, and (3) with no edge between Cell Size and Cell Shape.

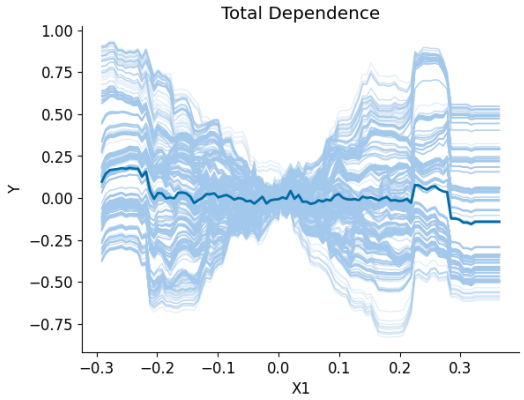

Figure 10: Individual counterfactual curves can show heterogeneous effects. In this example the relationship is positive for some individuals and negative for others with average effect of zero.

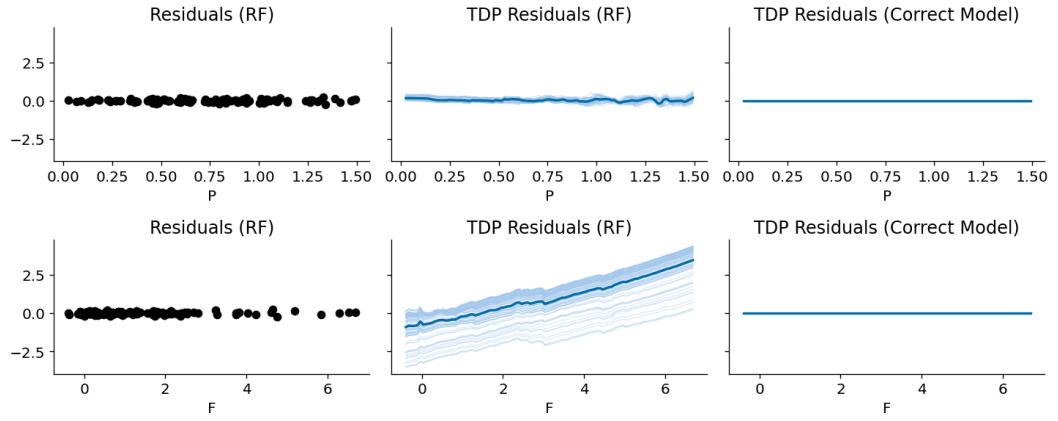

Figure 11: Regular versus CDP residuals for the example in Figure 1, plotted against feature P in the top row and F in the bottom row. Model multiplicity means two models can produce nearly the same predictions, with high accuracy, while using different functional relationships. Accuracy can only show if the model is "observationally correct," (left column) while CDPs can help determine if the model is also "causally correct" (middle vs right columns)

.

shift. Figure 11 demonstrates this with the salary example in Figure 1. Using the same RF model, residuals show no trend with respect to parental income P nor school funding F. However, a TDP over the residuals reveals the random forest model learns a causally incorrect functional form for $\hat{S}$, even if it fits the training data well. By comparison, only the causally correct model shows no trend in residuals with a CDP. This is an empirical verification of the correspondence between robustness to distribution shift and causal learning [46].

C Code and reproducibility

Predictive models were fit using `scikit-learn` [42]. CDP implementations make use of causal modeling functions in `dowhy` [53]. Figures were generated with `matplotlib` [20]. For causal structural learning we used the PC algorithm [55] implemented in Julia `CausalInference` [51]. We used the Python implementation of ALE plots in [22]. In experiments we used the Breast Cancer Wisconsin (Original) dataset [35, 59] and Sachs et al. [49] dataset. No specialized hardware is required to run the experiments as they are not computationally costly and can be reproduced on a personal computer. Our code to implement CDPs, run the experiments, and produce figures is available at this repository: `https://github.com/causalhypothesis/cdp-neurips/`

https://doi.org/10.52202/079017-3579

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction are that we developed a new type of model explanation plot that makes use of causal information about predictors. These claims accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed limitations in a subsection of the Discussion section, as well as throughout the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Most of the theoretical contributions of this paper consist in novel definitions. Our theoretical result establishing the equivalence of PDP+ICE with NDDP is proven by comparing their construction and holds without any additional assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Results are reproducible from the repository linked in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/079017-3579

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Results are reproducible from the repository linked in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The setup for each experiment is described in the paper and full details are given in supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The relevant methods for inference, that is, for non-parametric causal effects, are an area for future and ongoing research. There is not currently a consensus on which of the existing methods would be appropriate. Different types of CDPs may require completely different approaches for inference. In the present work we have not made claims that depend strongly on statistical significance. Future work using CDPs will have to choose an appropriate method of uncertainty estimation depending on the details of the given application, or, if none are available, develop new methods.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments are not computationally costly and can be reproduced on a personal computer.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research in this paper did not use human subjects or sensitive data. We have transparently communicated the limitations and potential impacts.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

https://doi.org/10.52202/079017-3579

Justification: Broader impacts are discussed, and are essentially those common to all model explanation/interpretation tools.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Code packages and datasets are cited and citations include licensing information.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code is documented in the linked repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects.

    Guidelines:

https://doi.org/10.52202/079017-3579

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.