Causal Dependence Plots for Interpretable Machine Learning

Joshua R. Loftus[1] **Lucius E. J. Bynum**[2] **Sakina Hansen**[1]

[1]Department of Statistics, London School of Economics, London, England, UK
[2]Center for Data Science, New York University, New York, NY, USA

1 INTRODUCTION

We propose Causal Dependence Plots (CDPs) to visualize relationships between input variables and a predicted outcome. Motivated by explaining or interpreting AI or machine learning models [1, 4, 5, 8], we focus on supervised learning, i.e. regression or classification, and specifically the model-agnostic or "black-box" setting. Model-agnostic interpretation methods are limited to observing how the model responds to variation in the inputs, and cannot access its internal structure.

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, the relationships revealed by focusing on one predictor at a time can be misleading. When varying one input variable, *we must make some choice about what values to use for the other inputs*. The Partial Dependence Plot (PDP) of Friedman [2] and Individual Conditional Expectation (ICE) plot from Goldstein et al. [3] are popular visual explanation methods. PDPs and ICE plots treat other predictors as independent of the one being plotted. This only captures the model dependence on each variable if predictors are independent and the model is additive [6]. Explanation methods may break—by ignoring—or respect existing statistical or causal dependencies between predictors.

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 [9, 14]. Such explanations could support spurious scientific hypotheses, 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 harm. For these reasons, *we care about the causal validity of model explanations*.

High level proposal. We use an auxiliary Explanatory Causal Model (ECM) to interpret or explain a given machine learning model. For each input predictor that we wish to explain, we use the ECM to determine how other inputs vary when that predictor is manipulated, rather than treating them as independent or fixed. We call the resulting plots Causal Dependence Plots or CDPs.

Pseudo-algorithm. To construct a CDP showing how $\hat{f}$ depends on x, a user specifies an explanatory causal model (ECM) M containing x and the other predictors, and an intervention $I(x)$ that manipulates x. The intervention $I(x)$ is chosen based on the specific explanation desired. An explanatory dataset D can be given or, if unavailable, generated by the ECM. For each observation i in D, and at each grid point x in the horizontal plot axis:

1. The ECM is used to simulate counterfactual values for all features of observation i under the intervention $I(x)$.
2. Counterfactual features are input into the prediction function $\hat{f}$, and the resulting counterfactual prediction is stored in an array indexed by (i, x).

Each observation in D then has an individual counterfactual prediction curve plotted against x. The empirical average of these curves is also plotted, and this is the main output of the CDP. The individual curves can be shown or suppressed as desired by the user. *The resulting CDP shows how the model's predictions causally depend on x when this predictor is varied by the intervention $I(x)$ in ECM M.*

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 $\hat{S} = \hat{f}(P, F)$. Blue curves show how $\hat{S}$ depends on P when P is causally manipulated *without* holding F constant, i.e. under the intervention $\text{do}(P = p)$. Orange curves show the dependence of $\hat{S}$ on P when F is

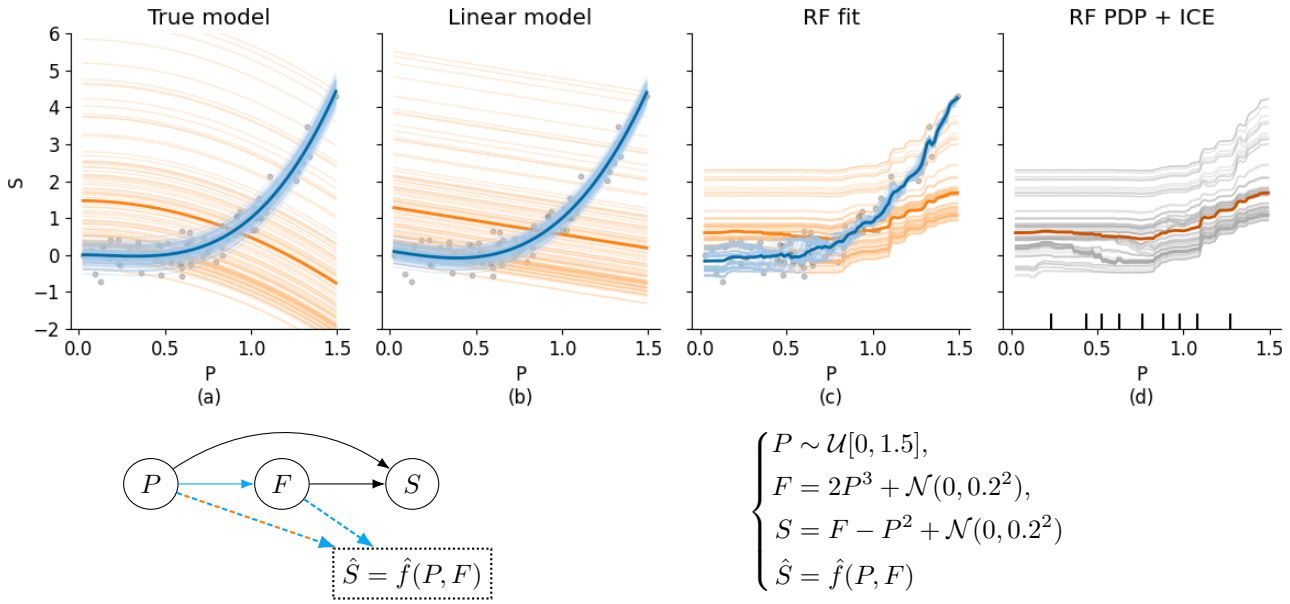

Figure 1: Motivating example. Causal Dependence Plots (top row) and the Explanatory Causal Model (bottom) for the motivating example. Total Dependence (TDP) is represented in blue and Natural Direct Dependence in orange. Panel (a) shows the relationships of the ECM. Counterfactual curves for individual points are shown as thin, light lines, with averages displayed as thick, dark lines. 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.

held constant at its observed value, and coincides exactly with a standard PDP. Several key takeaways are evident in Figure 1:

- *There can be qualitative differences between direct (or partial) dependence and total dependence*, a consequential fact when considering how interventions may change (predicted) outcomes. **An intervention which does *not* hold other predictors constant—arguably the canonical causal operation—can be shown by our TDP**. This is, to best of our knowledge, a novel contribution with high potential impact.
- *Our framework includes some existing model explanation plots like ICE and PDPs as special cases*. In panels (c-d), and later in Theorem B.2, 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.
- *Explanations of models can be qualitatively different from the underlying causal relationships*. For example, even a flexible model like 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. As another example, panel (b) shows that the total dependence of a linear model on a predictor can be non-linear because the mediator F depends non-linearly on P.

Discussion. The most important limitation for using CDPs is that they require specifying an ECM. Firstly, we assert this is an unavoidable requirement for any interpretation method to have causal relevance. Secondly, there are many potential applications based on different combinations of the predictive setting and choice of ECM, such as causal semi-supervised learning [12]. ECMs can be designed based on a particular desired explanation; make use of prior domain knowledge; and/or be learned and estimated from data using causal structural learning methods. Uncertainty about an ECM can be represented visually in the model explanation plot. We do not require a complete and correctly specified ECM to generate CDPs, they can be generated using only partial knowledge about predictors. Finally, CDPs can be useful for exploring model performance under covariate shift [13].

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[14] 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.

Causal Dependence Plots for Interpretable Machine Learning (Supplementary Material)

Joshua R. Loftus[1] **Lucius E. J. Bynum**[2] **Sakina Hansen**[1]

[1]Department of Statistics, London School of Economics, London, England, UK
[2]Center for Data Science, New York University, New York, NY, USA

A REAL DATA WITH DOMAIN KNOWLEDGE

An ECM may be constructed using domain expertise. Figure 2 shows an ECM and CDPs for the Sachs et al. [11] 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 [7]. 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.

B ALGORITHM DETAILS

For the following we assume predictor variables $\mathbf{X}$, an outcome of interest Y, and a black-box function $\hat{f}(x)$ with outputs that we may also denote $\hat{Y}$. A structural causal model $\mathcal{M}_{\mathbf{X}}$, either assumed or learned from data, specifies the causal relationships *only for the predictors* $\mathbf{X}$ and need not involve the outcome Y. Generating causal explanations for $\hat{f}$ involves performing abduction, action, and prediction with an ECM. In a large ECM graph we may suppress all arrows into $\hat{Y}$ except those from a single explanatory feature and its descendants. This is to simplify the display, as in Figure 2.

Definition B.1 (Explanatory Causal Model (ECM))**.** An ECM $\mathcal{M}$ augments the original SCM $\mathcal{M}_{\mathbf{X}}$ by including the predicted outcome $\hat{Y}$ as an additional variable with $\hat{f}$ as its structural equation.

Algorithm 1 Total Dependence Plot (TDP)
Inputs: $\mathcal{M}$ (ECM), $\hat{f}$ (black-box predictor), D (explanatory dataset), X_s (covariate of interest)

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

We often wish to decompose how much of the total effect of X on $\hat{Y}$ (or Y) is attributable to different pathways between the variables. This can be explored via direct dependence below.

[1]Following the discussion in [10] and follow-up ground truth DAG for the Sachs et al. [11] dataset in Figure 5 of [10], we choose the edge PIP3 $\rightarrow$ PIP2 in order to eliminate a would-be cycle.

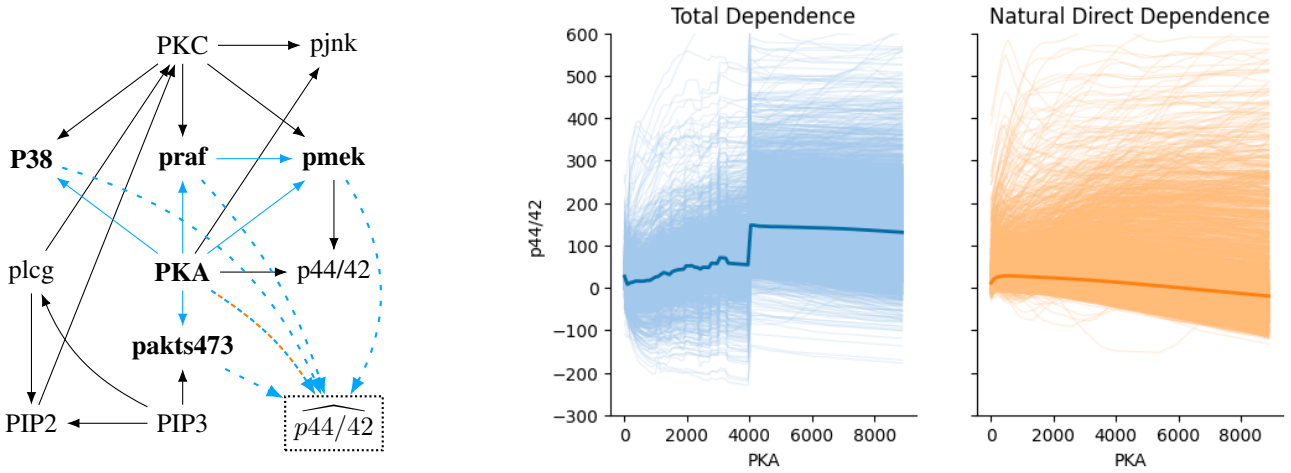

Figure 2: ECM for the Sachs et al. [11] dataset and corresponding CDPs for the effect of PKA on predicted p44/42. PKA and its descendants are bolded. While the NDDP (i.e. PDP + ICE) shows an overall decrease, the TDP shows an increase. *Conclusions depend strongly, qualitatively, on the specific interpretive question we ask, and causal modeling allows us to formulate questions precisely*.

Algorithm 2 Natural Direct Dependence Plot (NDDP)
Inputs: $\mathcal{M}$ (ECM), $\hat{f}$ (black-box predictor), D (explanatory dataset), X_s (covariate of interest)

Get the possible values of X_s and set to X
Set N to the number of observations in D
Initialize $N \times |X|$ matrix of estimates $\hat{Y}$
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 $D_{s \leftarrow x}$ entailed by $P^{\mathcal{M}|D;\text{do}(I,J)}$
  Set $\hat{Y}[:, x]$ to $\hat{f}(D_{s \leftarrow x})$
end for
Plot N lines $(X, \hat{Y}[i, :])$ {(Individual Counterfactuals)}
Plot average $(X, \sum_i \hat{Y}[i, :]/N)$ {(Causal Dependence)}

Comparing the construction of the NDDP to ICE curves and PDPs confirms what we observed in Figure 1(d).

Theorem B.2 (PDP + ICE = NDDP)**.** *When generating plots for the predictive model $\hat{f}$ using the dataset 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.*

Remark B.3. 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.