Practical do-Shapley Explanations with Estimand-Agnostic Causal Inference

Álvaro Parafita, Tomas Garriga,
Axel Brando, Francisco J. Cazorla

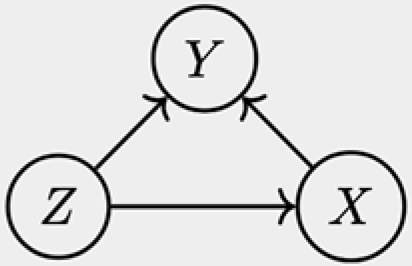
Introduction

- **Goal**: make causal explanations practical.
- **do-SHAP** (Jung et al. 2022): Shapley Values with *causal interventional* effects.

$$\phi_{\mathbf{x}}(X_k) = \sum_{\mathbf{S} \subseteq \mathbf{X} \setminus \{X_k\}} w(|\mathbf{S}|) \cdot (\nu_{\mathbf{x}}(\mathbf{S} \cup \{X_k\}) - \nu_{\mathbf{x}}(\mathbf{S})), \quad \nu_{\mathbf{x}}(\mathbf{S}) = \mathbb{E}[Y \mid do(\mathbf{S} = \mathbf{x}_{\mathbf{S}})]$$

- Two main **contributions**:
  1) A method to estimate $\nu_{\mathbf{x}}(\mathbf{S})$ in an automatable, practical way.
  2) A method to reduce the number of coalitions that need to be evaluated.

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Blöbaum, P., & Bareinboim, E. (2022, June). On measuring causal contributions via do-interventions. In *International Conference on Machine Learning* (pp. 10476-10501). PMLR.

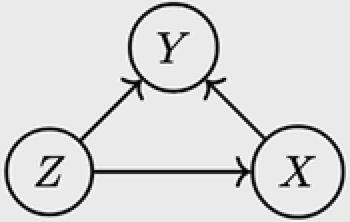


1) Why is do-SHAP impractical?

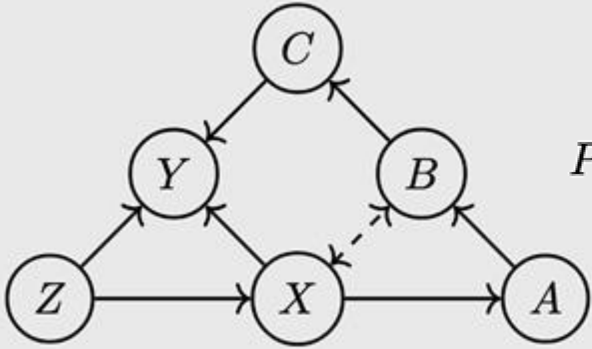
- By default, do-SHAP employs the **Estimand-Based approach**.
- Given a query (e.g., the effect of X on Y), compute an **estimand** of the query.

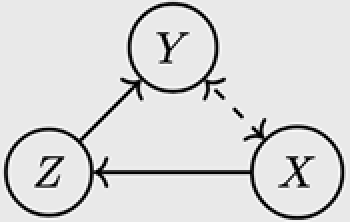
$$P_x(Y) = \mathbb{E}_Z\left[P(Y \mid x, Z)\right]$$

Estimand-Based approach



$$P_x(Y) = \mathbb{E}_Z\left[P(Y \mid x, Z)\right]$$





$$P_x(Y) = \mathbb{E}_{Z|x}\left[\mathbb{E}_X\left[P(Y \mid X, Z)\right]\right]$$

$$P_x(Y) = \mathbb{E}_{Z,C}\left[\frac{P(Y \mid x, Z, C)}{P(Z, C)} \cdot \mathbb{E}_{A|x}\left[\mathbb{E}_X\left[P(Z, C \mid X, A)\right]\right]\right]$$

Parafita, Á. (2022). *Causality without Estimands: from Causal Estimation to Black-Box Introspection* (Doctoral dissertation, Universitat de Barcelona (Spain)).

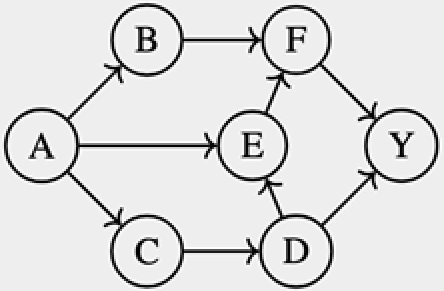
Estimands are impractical for do-SHAP

- **Problem**: do-SHAP needs to evaluate multiple queries, one per coalition:

$$\nu_{\mathbf{x}}(\mathbf{S}) = \mathbb{E}[Y \mid do(\mathbf{S} = \mathbf{x_S})]$$

- **Alternative:** Estimand-Agnostic Approach.
    - Train **a single Structural Causal Model**, following the known graph.
    - Use **general estimation procedures** to estimate any $\nu_{\mathbf{x}}(\mathbf{S})$.
- Guaranteed results as long as the queries are **identifiable.**

2) Reduce coalition evaluations



- Remove all **superfluous** nodes from any coalition.

$$\left.\begin{array}{l}\{A, B, C, E, F\},\ \{A, B, C, F\},\\ \{A, E, C, F\},\ \{B, E, C, F\},\\ \{A, C, F\},\ \{B, C, F\},\ \{E, C, F\}\end{array}\right\} = \{C, F\}$$

- **Cache**: only compute and store values for irreducible coalitions.

How to find the irreducible coalition?

Theorem. Given a topological order $<_{\mathcal{G}}$ in $\mathcal{G}$ and $\mathbf{S} \subseteq \mathbf{X}$, let $\mathbf{Z} := \{X \in \mathbf{S} \mid \mathbf{S}_{>_{\mathcal{G}} X} \in Fr_{\mathcal{G}}(X, Y)\}$, with $\mathbf{S}_{>_{\mathcal{G}} X} := \{Z \in \mathbf{S} \mid Z >_{\mathcal{G}} X\}$. Then $\nu(\mathbf{S}) = \nu(\mathbf{S} \setminus \mathbf{Z})$, and $\mathbf{S} \setminus \mathbf{Z}$ is irreducible.

Algorithm 1 Frontier-Reducibility Algorithm (FRA) – set version

Require: $\mathbf{S} \subseteq \mathbf{X}$, coalition.
Require: $\mathtt{Fr}$, a map: tuple[int] $\rightarrow$ bool.
Require: $\mathcal{G}$, causal graph.

procedure FRA($\mathbf{S}, \mathtt{Fr}; \mathcal{G}$)
  Sort($\mathbf{S}, <_{\mathcal{G}}$)
  $\mathbf{P} \leftarrow \varnothing$
  $\mathbf{Z} \leftarrow \varnothing$
  $k \leftarrow |\mathbf{S}|$
  while $k > 0$ **do**
    $X \leftarrow \mathbf{S}[k]$
    if $X \notin Pa_{\mathcal{G}}(Y)$ **then**
      $\mathbf{P'} \leftarrow \mathbf{P} \cap De_{\mathcal{G}}(X)$
      $\mathbf{T} \leftarrow (\mathbf{P'} \setminus \mathbf{Z}) \cup \{X\}$
      if $\mathbf{T} \notin \mathtt{Fr}$ **then**
        $\mathbf{C} \leftarrow \{X\}$
        while $\mathbf{C} \neq \varnothing$ and $Y \notin \mathbf{C}$ **do**
          $\mathbf{P'} \leftarrow \mathbf{P'} \cup \mathbf{C}$
          $\mathbf{C} \leftarrow \bigcup_{C \in \mathbf{C}} Ch_{\mathcal{G}}(C) \setminus \mathbf{P'}$
        end while
        $\mathtt{Fr}[\mathbf{T}] \leftarrow (C = \varnothing)$
      end if
      if $\mathtt{Fr}[\mathbf{T}]$ **then**
        $\mathbf{Z} \leftarrow \mathbf{Z} \cup \{X\}$
      end if
    end if
    $\mathbf{P} \leftarrow \mathbf{P} \cup \{X\}$
    $k \leftarrow k - 1$
  end while
  return $\mathbf{S} \setminus \mathbf{Z}$
end procedure

Conclusions

- **Estimand-Agnostic** Causal Inference as a **practical approach** for do-SHAP.
- **FRA** as an efficient algorithm to significantly **speed up do-SHAP**.

Questions?

- **Poster**: Wed 3 Dec, 11 a.m. – 2 p.m. PST
- **Email**: alvaro.parafita@bsc.es

arXiv

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

AI4S
BSC AI4Science Fellowships

AI4Debunk

The research leading to these results has received funding from the "Generacion D" initiative, Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by the European Union NextGenerationEU funds, through PRTR. It also was funded by the Horizon Europe Programme under the AI4DEBUNK Project (https://www.ai4debunk.eu), grant agreement num. 101135757. Additionally, this work has been partially supported by JDC2022-050313-I funded by MCIN/AEI/10.13039/501100011033al by the European Union NextGenerationEU/PRTR. We also gratefully acknowledge Novartis for sponsoring Tomàs Garriga's industrial PhD, and the Government of Catalonia's Industrial PhDs Plan for funding part of this research.