Unifying Feature-Based Explanations with Functional ANOVA and Cooperative Game Theory

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger

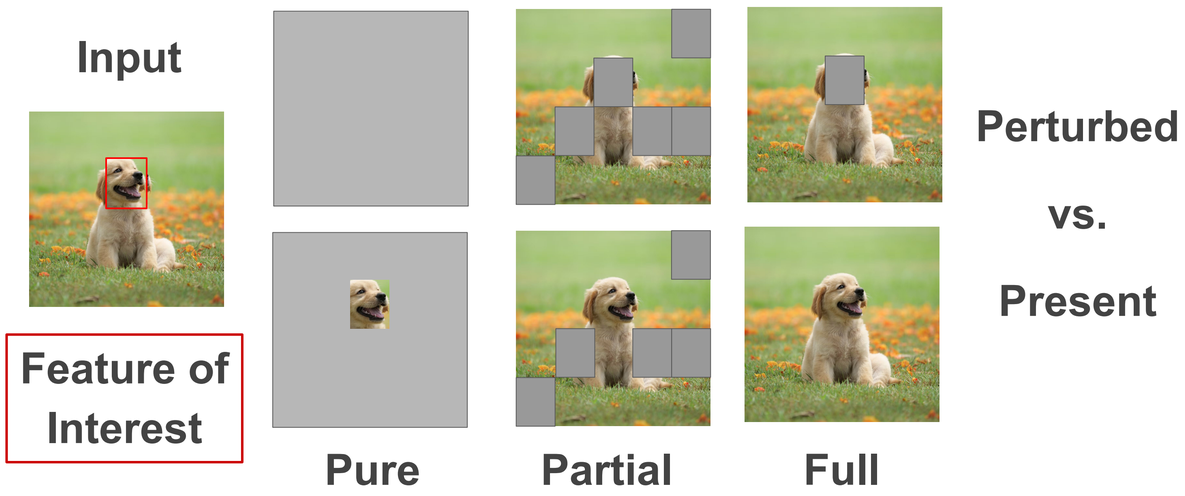
PROBLEM

Disagreement and fragmented landscape of feature-based explanations: Interpretation and comparison of feature-based explanations are unclear leading to confusion among practitioners.

Existing frameworks are limited by focusing on local [1] or global explanations [2,3], specific methods [4-6] or influences of single features [7].

COMPONENTS

Measuring Feature Influence by Perturbations



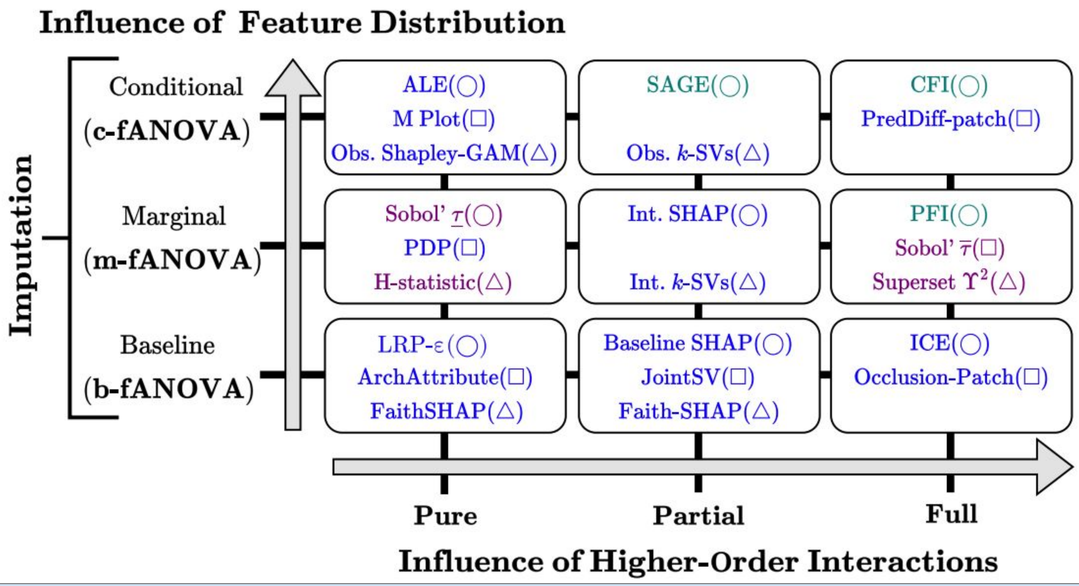
Imputation Methods for Perturbation

Value Function $F_S(x) := \int F(x)\, dP(x_{-S})$

Probability Distribution P determines Imputation

- **Baseline**: baseline value $F(x_S, b_{-S})$
- **Marginal**: background data $\mathbb{E}[F(x_S, X_{-S})]$
- **Conditional**: realistic data $\mathbb{E}[F(X) \mid X_S = x_S]$

UNIFYING FRAMEWORK



Explanation Game

Local	Model Prediction
Global	Model Performance (**risk**) Prediction Variance (**sensitivity**)

Explanation Type

One Feature (x_i)	○ Influence (**individual**)
Feature Groups $(x_i, x_j, \dots)$	□ Influence (**joint**) △ Synergies (**interaction**)

User Guide

1. **Explanation game: local** vs. **global**?
2. **Explanation Type: individuals**, groups (**joint**), or synergies (**interaction**)?
3. **Imputation: feature distribution** not (**baseline**), partially (**marginal**), or fully (**conditional**) captured?
4. **Higher-order interactions:** No (**pure**), **partial,** or **full** influence of higher-order interactions?

THEORETICAL FOUNDATION

Global Functional Decomposition via fANOVA

fANOVA effect $f_S(x) = F_S(x) - \sum_{T \subset S} f_T(x)$

yields $F(x) = f_\emptyset + \sum_{i=1}^{d} f_i(x_i) + \sum_{i \neq j} f_{ij}(x_i, x_j) \cdots = \sum_{S \subseteq D} f_S(x)$

Imputation Changes Influence of Feature Distribution

$F_{\text{lin}}(x) = \beta_1 x_1 + \beta_2 x_2 \qquad F_{\text{2int}}(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$

Table 1: Main and interaction fANOVA components of F_{lin} and F_{2int}, where $\bar{x}_i := x_i - \mu_i$

		b-fANOVA $f^{(b)}$	m-fANOVA $f^{(m)}$	c-fANOVA $f^{(c)}$
Main Effect f_i	F_{lin}	$\beta_i \cdot (x_i - b_i)$	$\beta_i \cdot \bar{x}_i$	$f^{(m)}_{\text{lin}} + \bar{x}_i \frac{\beta_j \sigma_{ij}}{\mathbb{V}[X_i]}$
	F_{2int}	$f^{(b)}_{\text{lin}} + \beta_{ij} b_j (x_i - b_i)$	$f^{(m)}_{\text{lin}} + \beta_{ij}\mu_j \bar{x}_i - \beta_{ij}\sigma_{ij}$	$f^{(c)}_{\text{lin}} + \beta_{ij}\bar{x}_i(\mu_j + \frac{\sigma_{ij} x_i}{\mathbb{V}[X_i]}) - \beta_{ij}\sigma_{ij}$
Interaction f_{ij}	F_{lin}	0	0	$-\sum_{\ell \in ij} \bar{x}_\ell \frac{\beta_{-\ell}\sigma_{-\ell,\ell}}{\mathbb{V}[X_\ell]}$
	F_{2int}	$\beta_{ij}(x_i - b_i)(x_j - b_j)$	$\beta_{ij}\bar{x}_i\bar{x}_j + \beta_{ij}\sigma_{ij}$	$f^{(m)}_{\text{2int}} + f^{(c)}_{\text{lin}} - \beta_{ij}\sigma_{ij}(\frac{\bar{x}_i x_i}{\mathbb{V}[X_i]} + \frac{\bar{x}_j x_j}{\mathbb{V}[X_j]})$

Feature Influence via Cooperative Game Theory

Explanation Game:

$\nu : 2^D \to \mathbb{R}$

Möbius Transform:

$m(S) = \sum_{T \subseteq S} (-1)^{s-t} \nu(T)$

Captures Properties of $F_S(x)$
- **Prediction** for local explanations
- **Variance or performance** for global explanations

Captures Properties of $f_S(x)$
- **fANOVA effect** for local explanations
- **Additive contribution** to variance or performance for global explanations

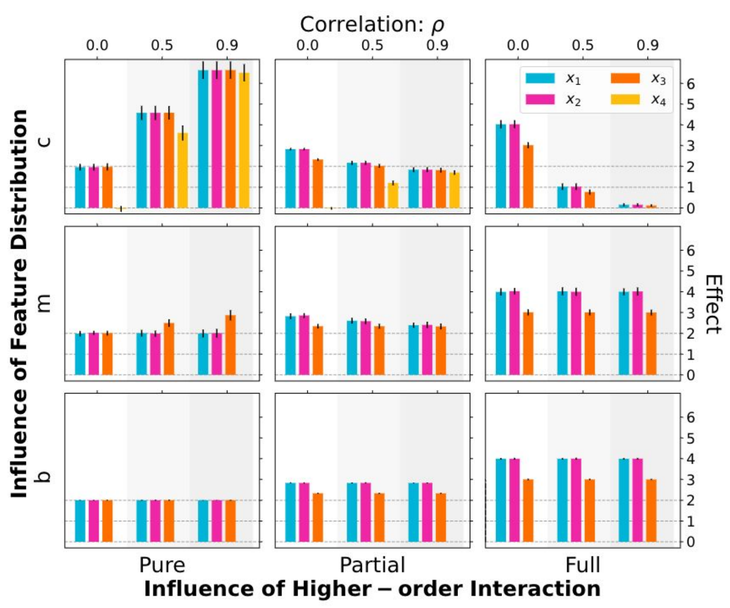
Pure, partial and full effects capture higher-order Möbius coefficients differently

Pure $(\phi^\emptyset)$	$m(i)$
Partial $(\phi^{\text{SV/GV/SI}})$	$m(i) + \sum_{T \supset i} \frac{m(T)}{t}$
Full (ϕ^+)	$m(i) + \sum_{T \supset i} m(T)$

EXPERIMENTS

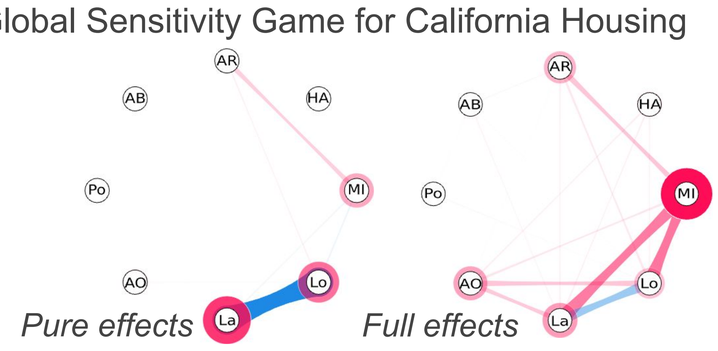
Synthetic Experiments

$F(x) = 2x_1 + 2x_2 + 2x_3 + x_1x_2 + x_1x_2x_3$

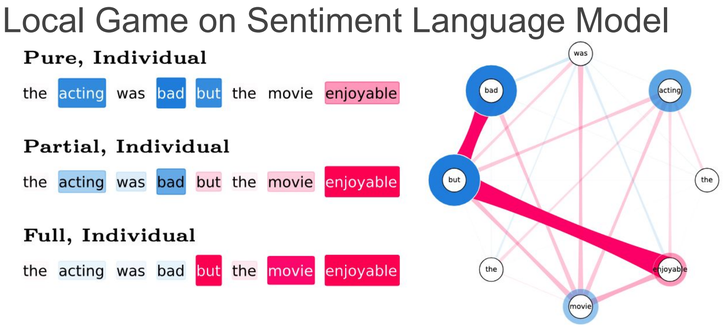


Real-world Applications

Global Sensitivity Game for California Housing



Local Game on Sentiment Language Model



Software & Reproducibility

Paper

Soon in **shapiq!**

SHAP-IQ

References

[1] Deng, H., Zou, N., Du, M., Chen, W., Feng, G., Yang, Z., Li, Z., and Zhang, Q. (2024). Unifying Fourteen Post-Hoc Attribution Methods With Taylor Interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7):4625–4640.
[2] Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding Global Feature Contributions With Additive Importance Measures. In NeurIPS'20, pages 17212–17223.
[3] Owen, A. B. (2013). Variance Components and Generalized Sobol' Indices. SIAM/ASA Journal on Uncertainty Quantification, 1(1):19–41.
[4] Lundstrom, D. and Razaviyayn, M. (2023). A unifying framework to the analysis of interaction methods using synergy functions. In ICML'23, pages 23005–23032.
[5] Bordt, S. and von Luxburg, U. (2023). From Shapley values to generalized additive models and back. In AISTATS'23, pages 709–745.
[6] Hiabu, M., Meyer, J. T., and Wright, M. N. (2023). Unifying local and global model explanations by functional decomposition of low dimensional structures. In AISTATS'23, pages 7040–7060.
[7] Covert, I., Lundberg, S., and Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. Journal of Machine Learning Research, 22(209):1–90.

UNIVERSITÄT BIELEFELD · LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN · Munich Center for Machine Learning · TRR 318 CONSTRUCTING EXPLAINABILITY · ATB Leibniz-Institut für Agrartechnik und Bioökonomie