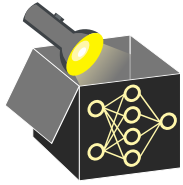
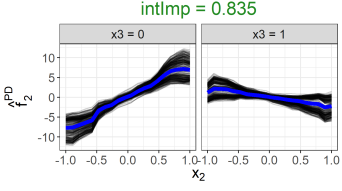
Interpretable Machine Learning



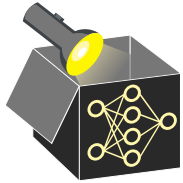
Regional Effects
Interaction importance



Learning goals

- Interaction quantification
- REPID interaction importance

INTERACTION QUANTIFICATION

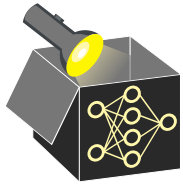


It's helpful to know not just how another feature changes the marginal effect of x_S but how strong that interaction is and want to rank features by it.

Approaches:

- **H-Statistics:** Variance of the deviation between the joint PDP and the sum of marginal PDPs (larger variance $\Rightarrow$ stronger interaction).
- **Greenwell's Interaction Index:** Difference between variance of the PDP and the mean variance of centered ICE curves for the same feature pair.
- **SHAP interaction index (▸ Herbinger 2022, ▸ Lundberg 2018):** Proportion of all two-way interactions with x_i to which the j-th feature contributes.

INTERACTION QUANTIFICATION



It's helpful to know not just how another feature changes the marginal effect of x_S but how strong that interaction is and want to rank features by it.

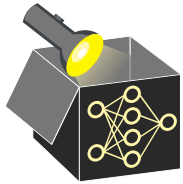
Approaches:

- **H-Statistics:** Variance of the deviation between the joint PDP and the sum of marginal PDPs (larger variance $\Rightarrow$ stronger interaction).
- **Greenwell's Interaction Index:** Difference between variance of the PDP and the mean variance of centered ICE curves for the same feature pair.
- **SHAP interaction index (** ▸ Herbinger 2022 **,** ▸ Lundberg 2018 **):** Proportion of all two-way interactions with x_i to which the j-th feature contributes.

Pitfalls:

- The values of H-Statistic and the Greenwell's Interaction Index are influenced by the main effects of the two regarded features.
- SHAP interaction index does not suffer from main effect problem. However, correlation between the two features can bias the interaction value. Same applies to the H-Statistic.

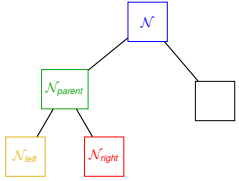
REPID INTERACTION IMPORTANCE



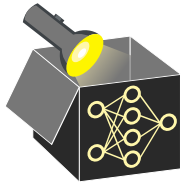
On parent node level (for $\mathcal{N}_{parent}$):

$$intImp(\mathcal{N}_{parent}) = \frac{\mathcal{R}(\mathcal{N}_{parent}) - (\mathcal{R}(\mathcal{N}_{left}) + \mathcal{R}(\mathcal{N}_{right}))}{\mathcal{R}(\mathcal{N})}$$

Interpretation: Reduction of ICE curve variance after one split of $\mathcal{N}_{parent}$ into $\mathcal{N}_{left}$ and $\mathcal{N}_{right}$ relative to the ICE curve variance in the root node $\mathcal{N}$.



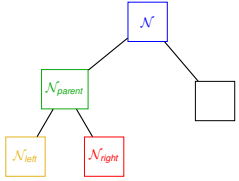
REPID INTERACTION IMPORTANCE

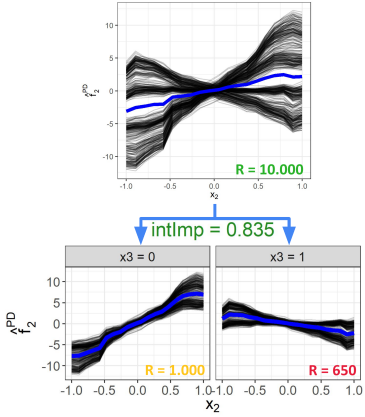


On parent node level (for $\mathcal{N}_{parent}$):

$$intImp(\mathcal{N}_{parent}) = \frac{\mathcal{R}(\mathcal{N}_{parent}) - (\mathcal{R}(\mathcal{N}_{left}) + \mathcal{R}(\mathcal{N}_{right}))}{\mathcal{R}(\mathcal{N})}$$

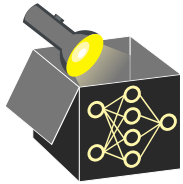
Interpretation: Reduction of ICE curve variance after one split of $\mathcal{N}_{parent}$ into $\mathcal{N}_{left}$ and $\mathcal{N}_{right}$ relative to the ICE curve variance in the root node $\mathcal{N}$.





Split reduces 83.5% of variance.

REPID INTERACTION IMPORTANCE



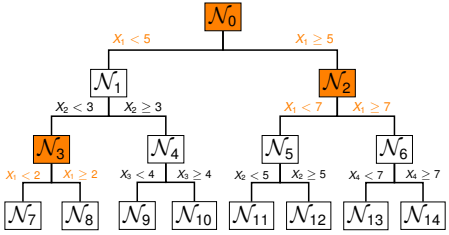
On feature level (for x_j):

$$intImp_j = \sum_{i \in \mathcal{B}_j} intImp(\mathcal{N}_i)$$

where $\mathcal{B}_j$ indexes parent nodes split by x_j.

Interpretation: Overall reduction of ICE curve variance due to splits by X_j (in %).

Example: For $X_1 \Rightarrow \mathcal{B}_1 = \{0, 2, 3\}$



REPID INTERACTION IMPORTANCE

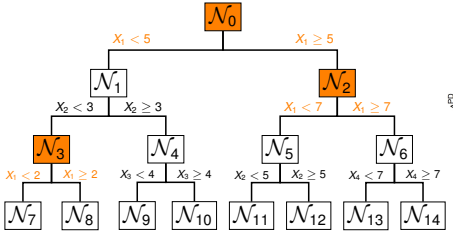
On feature level (for x_j):

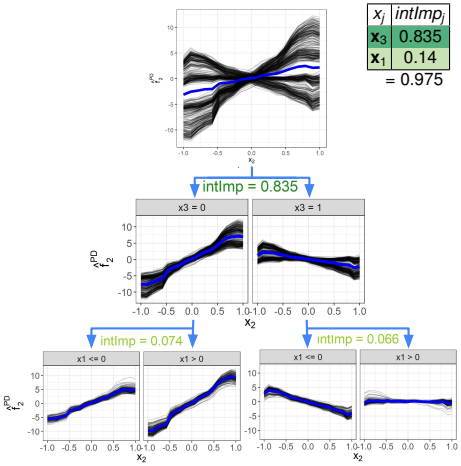
$$intImp_j = \sum_{i \in \mathcal{B}_j} intImp(\mathcal{N}_i)$$

where $\mathcal{B}_j$ indexes parent nodes split by x_j.

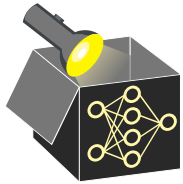
Interpretation: Overall reduction of ICE curve variance due to splits by X_j (in %).

Example: For $X_1 \Rightarrow \mathcal{B}_1 = \{0, 2, 3\}$



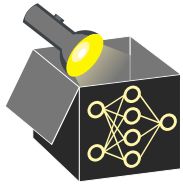


x_j	$intImp_j$
x_3	0.835
x_1	0.14
= 0.975	



Note: *intImp* can also be used as a stopping criterion.

OUTPERFORMING SOTA



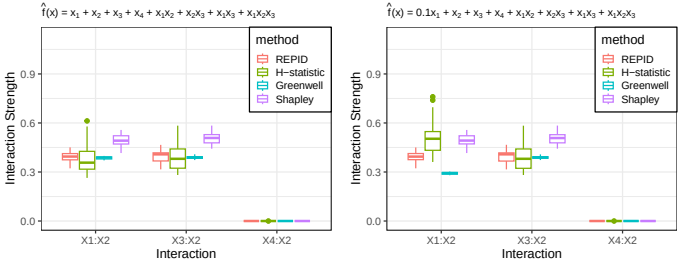
Simulation setting

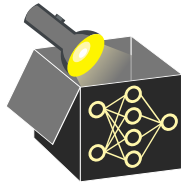
- Draw 1000 i.i.d. samples from $X_1, \ldots, X_4 \sim \mathcal{U}(-1, 1)$
- True underlying function:
  $f(\mathbf{x}) = \sum_{j=1}^{4} \mathbf{x}_j + \mathbf{x}_1\mathbf{x}_2 + \mathbf{x}_2\mathbf{x}_3 + \mathbf{x}_1\mathbf{x}_3 + \mathbf{x}_1\mathbf{x}_2\mathbf{x}_3 + \epsilon$
- Fit a correctly specified linear model (interactions with $\mathbf{x}_4$ are excluded)
- 30 reps, measure interaction strength between $\mathbf{x}_2$ and all other 3 features

Which methods are sensitive to changes in main effect sizes or feature correlations?

Pitfall	REPID	H-Statistic	Greenwell	SHAP
sensitive to changes of main effect	No	Yes	Yes	No
sensitive to changes of correlation between $\mathbf{x}_j$ and other features	No	Yes	No	Yes

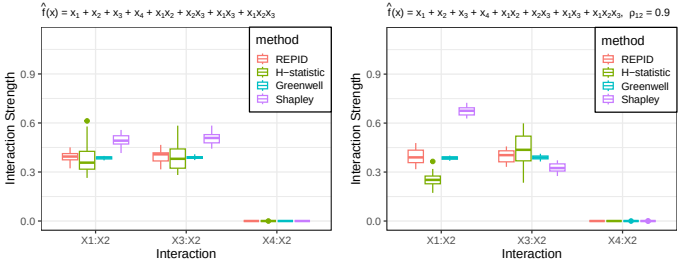
OUTPERFORMING SOTA

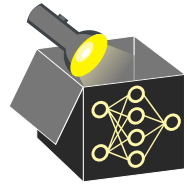




- **Left (initial setting)**: Interaction strength of $\mathbf{x}_1$:$\mathbf{x}_2$ and $\mathbf{x}_3$:$\mathbf{x}_2$ similar; $\mathbf{x}_4$:$\mathbf{x}_2$ no interaction
- **Right**: Set main effect $\beta_1 = 0.1$
  - **Expectation**: Interaction strengths should not change
  - **Fail**: H-statistic ($\mathbf{x}_1$:$\mathbf{x}_2$ > $\mathbf{x}_3$:$\mathbf{x}_2$) and Greenwell ($\mathbf{x}_1$:$\mathbf{x}_2$ < $\mathbf{x}_3$:$\mathbf{x}_2$)

OUTPERFORMING SOTA



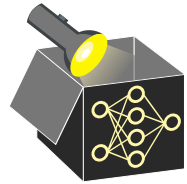


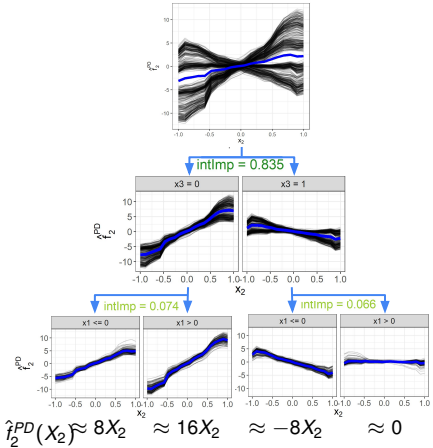
- **Left (initial setting)**: Interaction strength of $\mathbf{x}_1$:$\mathbf{x}_2$ and $\mathbf{x}_3$:$\mathbf{x}_2$ similar; $\mathbf{x}_4$:$\mathbf{x}_2$ no interaction
- **Right**: Increase correlation $\rho(\mathbf{x}_1, \mathbf{x}_2) = 0.9$
  - **Expectation**: Interaction strengths should not change
  - **Fail**: H-statistic ($\mathbf{x}_1$:$\mathbf{x}_2$ < $\mathbf{x}_3$:$\mathbf{x}_2$) and Shapley ($\mathbf{x}_1$:$\mathbf{x}_2$ > $\mathbf{x}_3$:$\mathbf{x}_2$)

$\rightarrow$ **REPID is the only method which always leads to correct rankings for these settings**

LIMITATIONS OF REPID

1) Restricted to one feature of interest
$\Rightarrow$ Different regions for different features





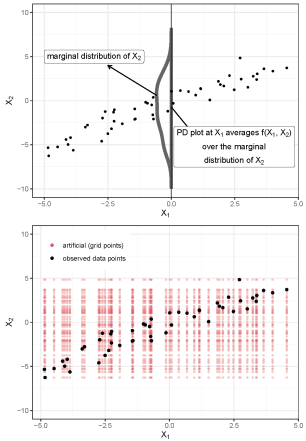
$\hat{f}_2^{PD}(X_2) \approx 8X_2 \quad \approx 16X_2 \quad \approx -8X_2 \quad \approx 0$

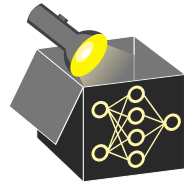
LIMITATIONS OF REPID

1) Restricted to one feature of interest
$\Rightarrow$ Different regions for different features

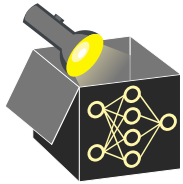
2) Restricted to PD (global) and ICE (local) as feature effect methods
$\Rightarrow$ Inherits extrapolation problem (unlikely combinations of feature values)

$\rightsquigarrow$ Follow-up GADGET [under review]





CONCLUSION

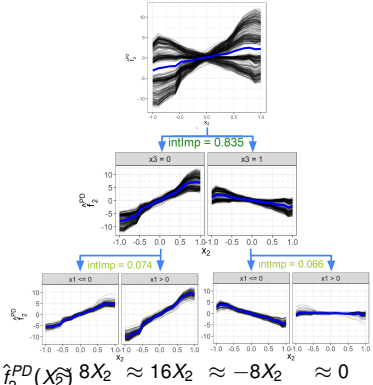


Summary of Contributions (REPID):

- Regional effects in interpretable regions
- Additive decomp. of feature effect
- Quantify feature interactions
- Outperforms SOTA interaction indices

Summary of Contributions (GADGET):

- Unique regions for multiple features
- Additive decomp. of prediction function
- Extension to ALE and Shapley Dependence
- Test to identify significant interactions



Further Directions:

Pruning, GADGET as a predictor, comparing regions across models, efficient implementation, more efficient testing and splitting approach, . . .