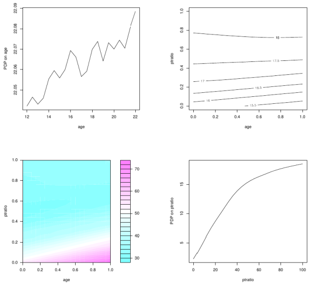
Interpretable Machine Learning

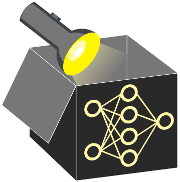
Functional Decompositions
Further Methods



Learning goals

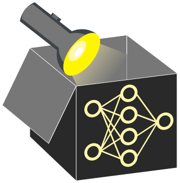
- Limitations of classical fANOVA
- Alternatives: Generalized fANOVA and ALE
- Advantages and relevance of functional decompositions

LIMITATIONS OF CLASSICAL FANOVA



- Standard fANOVA builds on PD-functions
- *Remember:* Problems of PDPs for **correlated** / **dependent features**
- Here: Dependent features $\Longrightarrow$ Standard fANOVA does NOT fulfill vanishing conditions

LIMITATIONS OF CLASSICAL FANOVA



- Standard fANOVA builds on PD-functions
- *Remember:* Problems of PDPs for **correlated** / **dependent features**
- Here: Dependent features $\Longrightarrow$ Standard fANOVA does NOT fulfill vanishing conditions

Example

Assume dependency $2x_1^2 = x_2$ and

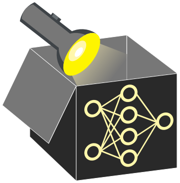
$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_2x_3 + 1.$$

$\rightsquigarrow$ Following two decompositions would both "make sense":

$$\hat{f}(x_1, x_2, x_3) = \underbrace{1}_{g_\emptyset} + \underbrace{(-2x_1)}_{g_1(x_1)} + \underbrace{(-2\sin(x_3))}_{g_3(x_3)} + \underbrace{|x_1|x_2}_{g_{1,2}(x_1,x_2)} + \underbrace{0.5x_2x_3}_{g_{2,3}(x_2,x_3)}$$

$$\hat{f}(x_1, x_2, x_3) = \underbrace{1}_{g_\emptyset} + \underbrace{(-2x_1 + 2|x_1|^3)}_{g_1(x_1)} + \underbrace{(-2\sin(x_3))}_{g_3(x_3)} + \underbrace{x_1^2x_3}_{g_{2,3}(x_1,x_3)}$$

LIMITATIONS OF CLASSICAL FANOVA



- Standard fANOVA builds on PD-functions
- *Remember:* Problems of PDPs for **correlated / dependent features**
- Here: Dependent features $\Longrightarrow$ Standard fANOVA does NOT fulfill vanishing conditions

Example

Assume dependency $2x_1^2 = x_2$ and

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_2x_3 + 1.$$

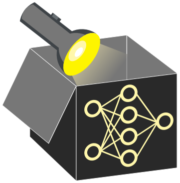
$\rightsquigarrow$ Following two decompositions would both “make sense”:

$$\hat{f}(x_1, x_2, x_3) = \underbrace{1}_{g_\emptyset} + \underbrace{(-2x_1)}_{g_1(x_1)} + \underbrace{(-2\sin(x_3))}_{g_3(x_3)} + \underbrace{|x_1|x_2}_{g_{1,2}(x_1,x_2)} + \underbrace{0.5x_2x_3}_{g_{2,3}(x_2,x_3)}$$

$$\hat{f}(x_1, x_2, x_3) = \underbrace{1}_{g_\emptyset} + \underbrace{(-2x_1 + 2|x_1|^3)}_{g_1(x_1)} + \underbrace{(-2\sin(x_3))}_{g_3(x_3)} + \underbrace{x_1^2x_3}_{g_{2,3}(x_1,x_3)}$$

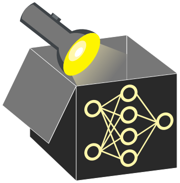
$\rightarrow$ Extreme example, but again: Problem of definition

ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA



- Algorithm proposed by ▶ "Hooker" 2007
- Generalizes standard fANOVA to situations with dependent features

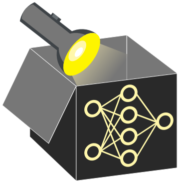
ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA



- Algorithm proposed by ▸ "Hooker" 2007
- Generalizes standard fANOVA to situations with dependent features
- Showed: Generalized fANOVA is solution to so-called "relaxed vanishing conditions"
  (i.e., weaker form of vanishing condition)
- "Relaxed vanishing conditions" do not imply orthogonality, but "hierarchical orthogonality":

$$\mathbb{E}_{\mathbf{x}}\left[g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)\right] = 0 \quad \forall V \subsetneq S$$

ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA



- Algorithm proposed by ▸ "Hooker" 2007
- Generalizes standard fANOVA to situations with dependent features
- Showed: Generalized fANOVA is solution to so-called "relaxed vanishing conditions"
  (i.e., weaker form of vanishing condition)
- "Relaxed vanishing conditions" do not imply orthogonality, but "hierarchical orthogonality":

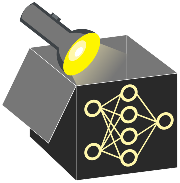
$$\mathbb{E}_{\mathbf{x}}\left[g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)\right] = 0 \quad \forall V \subsetneq S$$

$\rightsquigarrow$ Only components are orthogonal where $g_V(\mathbf{x}_V)$ is "lower in hierarchy" than $g_S(\mathbf{x}_S)$

$\Longrightarrow$ Generalized fANOVA provides functional decomp. for arbitrary settings

- **Advantage:** Also provides a variance decomposition

ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA



- Algorithm proposed by ▸ "Hooker" 2007
- Generalizes standard fANOVA to situations with dependent features
- Showed: Generalized fANOVA is solution to so-called "relaxed vanishing conditions"
  (i.e., weaker form of vanishing condition)
- "Relaxed vanishing conditions" do not imply orthogonality, but "hierarchical orthogonality":

$$\mathbb{E}_{\mathbf{x}}\left[g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)\right] = 0 \quad \forall V \subsetneq S$$

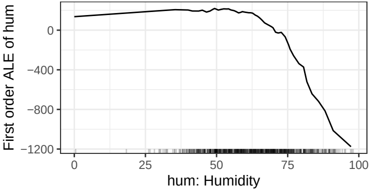
  $\rightsquigarrow$ Only components are orthogonal where $g_V(\mathbf{x}_V)$ is "lower in hierarchy" than $g_S(\mathbf{x}_S)$

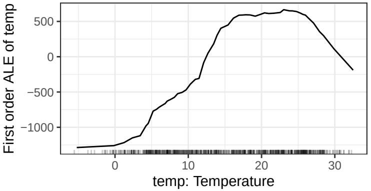
$\Longrightarrow$ Generalized fANOVA provides functional decomp. for arbitrary settings

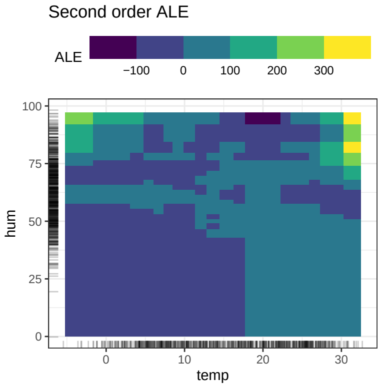
- **Advantage:** Also provides a variance decomposition
- **Problems:**
- Difficult to estimate, involves manual choice of a "weight function"
- Computationally very costly

REVISITING ALE PLOTS

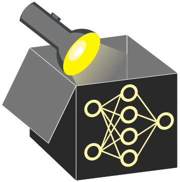
$$\hat{\tilde{f}}_{S,ALE}(x) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i:\ x_S^{(i)} \in\ ]z_{k-1,S}, z_{k,S}]} \left[\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)}) \right]$$







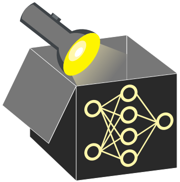
ALE DECOMPOSITION



- One can define ALE plots for arbitrary many variables (similar to PDPs vs. PD-functions)

$\rightarrow$ Gives full functional decomposition of ALE plots

ALE DECOMPOSITION



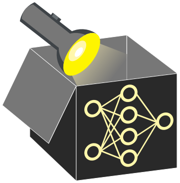
- One can define ALE plots for arbitrary many variables (similar to PDPs vs. PD-functions)

$\rightarrow$ Gives full functional decomposition of ALE plots

- **Advantages:** Handle dependencies well + computationally fast
- Constraints / orthogonality properties more complicated

$\Longrightarrow$ ALE decomp. theoretically more involved, but good alternative in practice

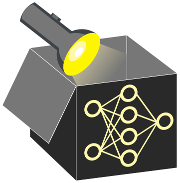
CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



- If computed, offer many insights into a model / function, i.p. high-dim.

$\rightarrow$ Complete analysis of all interactions

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?

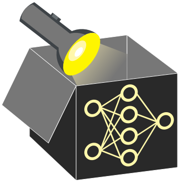


- If computed, offer many insights into a model / function, i.p. high-dim.

$\rightarrow$ Complete analysis of all interactions

- Very important theoretical concept:
  - Theoretical framework for general definition of interact.-s (H-statistic)
  - Theoretical background for many IML methods: GAMs and EBMs, ICE, PDPs and PD-functions, ALE plots, Shapley values, Feature importance methods (see later)

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



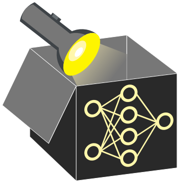
- If computed, offer many insights into a model / function, i.p. high-dim.

$\rightarrow$ Complete analysis of all interactions

- Very important theoretical concept:
  - Theoretical framework for general definition of interact.-s (H-statistic)
  - Theoretical background for many IML methods: GAMs and EBMs, ICE, PDPs and PD-functions, ALE plots, Shapley values, Feature importance methods (see later)
- In practice often infeasible (2^p components for p features)

$\Longrightarrow$ Often only sparse decompositions feasible (e.g. EBMs)

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



- If computed, offer many insights into a model / function, i.p. high-dim.

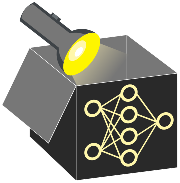
$\rightarrow$ Complete analysis of all interactions

- Very important theoretical concept:
  - Theoretical framework for general definition of interact.-s (H-statistic)
  - Theoretical background for many IML methods: GAMs and EBMs, ICE, PDPs and PD-functions, ALE plots, Shapley values, Feature importance methods (see later)
- In practice often infeasible (2^p components for p features)

$\Longrightarrow$ Often only sparse decompositions feasible (e.g. EBMs)

- All single methods have disadvantages:
  - Standard fANOVA: Only independent features + compute intensive
  - Generalized fANOVA: Even more computational intensive, eventually infeasible
  - ALE: No variance decomposition

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



- If computed, offer many insights into a model / function, i.p. high-dim.

$\rightarrow$ Complete analysis of all interactions

- Very important theoretical concept:
  - Theoretical framework for general definition of interact.-s (H-statistic)
  - Theoretical background for many IML methods: GAMs and EBMs, ICE, PDPs and PD-functions, ALE plots, Shapley values, Feature importance methods (see later)
- In practice often infeasible (2^p components for p features)

$\Longrightarrow$ Often only sparse decompositions feasible (e.g. EBMs)

- All single methods have disadvantages:
  - Standard fANOVA: Only independent features + compute intensive
  - Generalized fANOVA: Even more computational intensive, eventually infeasible
  - ALE: No variance decomposition

Overall: Very important concept and theoretical background, explains idea behind many other methods