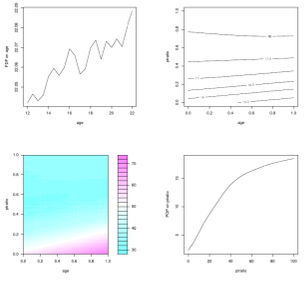
Interpretable Machine Learning

Introduction to Functional Decomposition

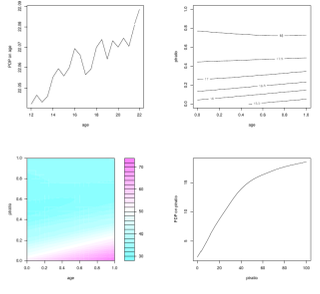


Learning goals

- Basic idea of additive functional decompositions
- Motivation and usefulness of functional decompositions
- Difficulty of obtaining or even defining a functional decomposition
- Several examples

Interpretable Machine Learning

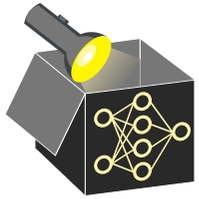
Functional Decompositions Introduction



Learning goals

- Basic idea of additive functional decompositions
- Motivation and usefulness of functional decompositions
- Difficulty of obtaining or even defining a functional decomposition
- Several examples

PRELIMINARIES

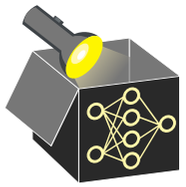


Recap: Interactions

- Interactions between features: Effect of one feature on the prediction output depends on (one or more) other features
- Definition: Features x_j and x_k are considered to interact, if

$$\mathbb{E}\left[\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}\right)^2\right] > 0$$

PRELIMINARIES

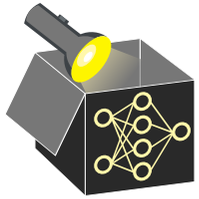


Recap: Interactions

- Interactions between features: Effect of one feature on the prediction output depends on (one or more) other features
- Definition: Features x_j and x_k are considered to interact, if

$$\mathbb{E}\left[\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}\right)^2\right] > 0$$

PRELIMINARIES



Recap: Interactions

- Interactions between features: Effect of one feature on the prediction output depends on (one or more) other features
- Definition: Features x_j and x_k are considered to interact, if

$$\mathbb{E}\left[\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}\right)^2\right] > 0$$

Recap: GAMs

- Decomposition into only main effects
- Do not contain any interactions

$$\hat{f}(\mathbf{x}) = \theta_0 + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

PRELIMINARIES

Recap: Interactions

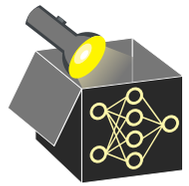
- Interactions between features: Effect of one feature on the prediction output depends on (one or more) other features
- Definition: Features x_j and x_k are considered to interact, if

$$\mathbb{E}\left[\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}\right)^2\right] > 0$$

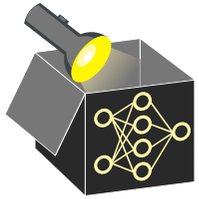
Recap: GAMs

- Decomposition into only main effects
- Do not contain any interactions

$$\hat{f}(\mathbf{x}) = \theta_0 + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$



FIRST EXAMPLE: ADDITIVE DECOMPOSITION



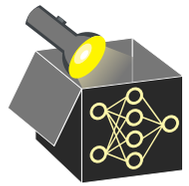
Example

Consider

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- Idea: Additive decomposition depending on which features used:

FIRST EXAMPLE: ADDITIVE DECOMPOSITION



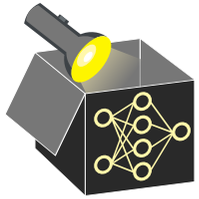
Example

Consider

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- Idea: Additive decomposition depending on which features used:

FIRST EXAMPLE: ADDITIVE DECOMPOSITION



Example

Consider

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- Idea: Additive decomposition depending on which features used:

$$\begin{aligned} g_\emptyset(x_1, x_2) &= 4 && \text{Part depending on no features at all (intercept)} \\ \left.\begin{aligned} g_1(x_1, x_2) &= 2x_1 \\ g_2(x_1, x_2) &= 0.3e^{x_2} \end{aligned}\right\} && \text{Parts depending on a single feature (main effects)} \quad (1) \\ g_{1,2}(x_1, x_2) &= |x_1|x_2 && \text{Part depending on both features (interaction)} \end{aligned}$$

$\rightsquigarrow$ single terms with immediate interpretation, full understanding of the model

FIRST EXAMPLE: ADDITIVE DECOMPOSITION

Example

Consider

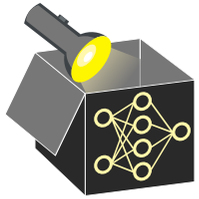
$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- Idea: Additive decomposition depending on which features used:

$$\begin{aligned} g_\emptyset(x_1, x_2) &= 4 && \text{Part depending on no features at all (intercept)} \\ \left.\begin{aligned} g_1(x_1, x_2) &= 2x_1 \\ g_2(x_1, x_2) &= 0.3e^{x_2} \end{aligned}\right\} && \text{Parts depending on a single feature (main effects)} \\ g_{1,2}(x_1, x_2) &= |x_1|x_2 && \text{Part depending on both features (interaction)} \quad (1) \end{aligned}$$

$\rightsquigarrow$ Single terms with immediate interpretation, full model understanding

FIRST EXAMPLE: ADDITIVE DECOMPOSITION



Example

Consider

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

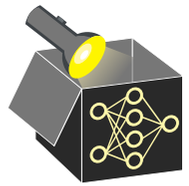
- Idea: Additive decomposition depending on which features used:

$$\begin{aligned} g_\emptyset(x_1, x_2) &= 4 && \text{Part depending on no features at all (intercept)} \\ \left.\begin{aligned} g_1(x_1, x_2) &= 2x_1 \\ g_2(x_1, x_2) &= 0.3e^{x_2} \end{aligned}\right\} && \text{Parts depending on a single feature (main effects)} \quad (1) \\ g_{1,2}(x_1, x_2) &= |x_1|x_2 && \text{Part depending on both features (interaction)} \end{aligned}$$

$\rightsquigarrow$ single terms with immediate interpretation, full understanding of the model

$\leftrightarrow$ Not possible with effects of single features (e.g. PDPs) or GAM surrogate model (miss interaction part)

FIRST EXAMPLE: ADDITIVE DECOMPOSITION



Example

Consider

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- Idea: Additive decomposition depending on which features used:

$$\begin{aligned} g_\emptyset(x_1, x_2) &= 4 && \text{Part depending on no features at all (intercept)} \\ \left.\begin{aligned} g_1(x_1, x_2) &= 2x_1 \\ g_2(x_1, x_2) &= 0.3e^{x_2} \end{aligned}\right\} && \text{Parts depending on a single feature (main effects)} \\ g_{1,2}(x_1, x_2) &= |x_1|x_2 && \text{Part depending on both features (interaction)} \quad (1) \end{aligned}$$

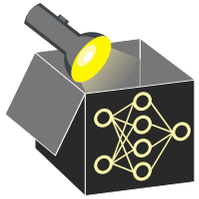
$\rightsquigarrow$ Single terms with immediate interpretation, full model understanding

$\leftrightarrow$ Not possible with effects of single features (e.g. PDPs) or GAM surrogate model (miss interaction part)

ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION

Goal in general: Given a black-box model $\hat{f} : \mathbb{R}^2 \to \mathbb{R}$, find a decomposition

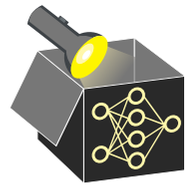
$$\hat{f}(x_1, x_2) = g_\emptyset + g_1(x_1) + g_2(x_2) + g_{1,2}(x_1, x_2) \tag{2}$$



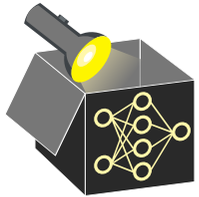
ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION

Goal in general: Given a black-box model $\hat{f} : \mathbb{R}^2 \to \mathbb{R}$, find a decomposition

$$\hat{f}(x_1, x_2) = g_\emptyset + g_1(x_1) + g_2(x_2) + g_{1,2}(x_1, x_2) \tag{2}$$



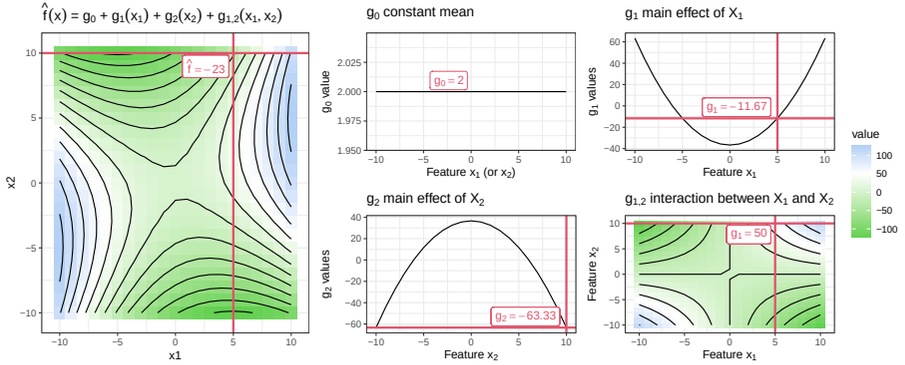
ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION



Goal in general: Given a black-box model $\hat{f} : \mathbb{R}^2 \to \mathbb{R}$, find a decomposition

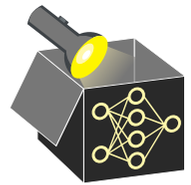
$$\hat{f}(x_1, x_2) = g_\emptyset + g_1(x_1) + g_2(x_2) + g_{1,2}(x_1, x_2) \tag{2}$$

Example



⇝ More details on this example later

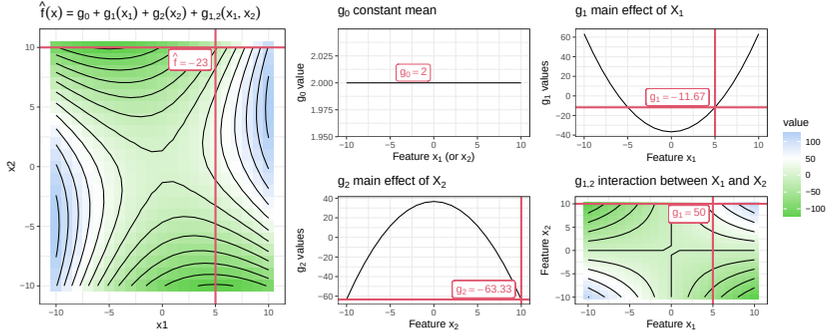
ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION



Goal in general: Given a black-box model $\hat{f} : \mathbb{R}^2 \to \mathbb{R}$, find a decomposition

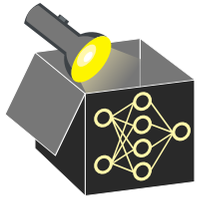
$$\hat{f}(x_1, x_2) = g_\emptyset + g_1(x_1) + g_2(x_2) + g_{1,2}(x_1, x_2) \tag{2}$$

Example



⇝ More details on this example later

ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION

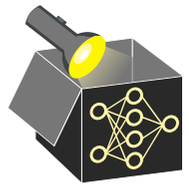


Example

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

Again, read additive decomposition from formula:

ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION

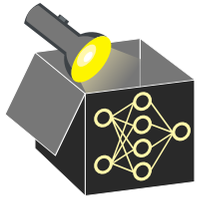


Example

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

Again, read additive decomposition from formula:

ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION



Example

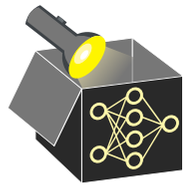
$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

Again, read additive decomposition from formula:

$$\begin{aligned}
&g_\emptyset(x_1, x_2, x_3) = 1 && \text{constant part, no effects}\\
&\left.\begin{aligned} g_1(x_1, x_2, x_3) &= -2x_1\\ g_2(x_1, x_2, x_3) &= 0\\ g_3(x_1, x_2, x_3) &= -2\sin(x_3)\end{aligned}\right\} && \text{main effects, no interactions}\\
&\left.\begin{aligned} g_{1,2}(x_1, x_2, x_3) &= |x_1|x_2\\ g_{1,3}(x_1, x_2, x_3) &= 0\\ g_{2,3}(x_1, x_2, x_3) &= -\sin(x_2x_3)\end{aligned}\right\} && \text{2-way interactions (depending on 2 features)}\\
&g_{1,2,3}(x_1, x_2, x_3) = 0.5x_1x_2x_3 && \text{3-way interactions}
\end{aligned} \tag{3}$$

$\Rightarrow$ 8 components in total, but some empty $\rightsquigarrow$ Certain interactions not present

ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION



Example

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

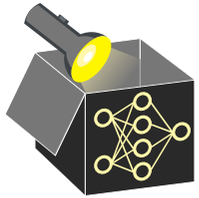
Again, read additive decomposition from formula:

$$\begin{aligned}
&g_\emptyset(x_1, x_2, x_3) = 1 && \text{constant part, no effects}\\
&\left.\begin{aligned} g_1(x_1, x_2, x_3) &= -2x_1\\ g_2(x_1, x_2, x_3) &= 0\\ g_3(x_1, x_2, x_3) &= -2\sin(x_3)\end{aligned}\right\} && \text{main effects, no interactions}\\
&\left.\begin{aligned} g_{1,2}(x_1, x_2, x_3) &= |x_1|x_2\\ g_{1,3}(x_1, x_2, x_3) &= 0\\ g_{2,3}(x_1, x_2, x_3) &= -\sin(x_2x_3)\end{aligned}\right\} && \text{2-way interactions (depending on 2 features)}\\
&g_{1,2,3}(x_1, x_2, x_3) = 0.5x_1x_2x_3 && \text{3-way interactions}
\end{aligned} \tag{3}$$

$\Rightarrow$ 8 components in total, but some empty $\rightsquigarrow$ Certain interactions not present

GENERAL FORM OF FUNCTIONAL DECOMPOSITION

Li and Rabitz (2011) | Chastaing et al. (2012)



Definition

Functional decomposition: Additive decomposition of a function $\hat{f} : \mathbb{R}^p \mapsto \mathbb{R}$ into a sum of components of different dimensions w.r.t. inputs $x_1, \ldots, x_p$:

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \sum_{S \subseteq \{1,\ldots,p\}} g_S(\mathbf{x}_S) \\ &= g_\emptyset + g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) + \\ &\quad g_{1,2}(x_1, x_2) + \cdots + g_{p-1,p}(x_{p-1}, x_p) + \cdots + \\ &\quad g_{1,2,3}(x_1, x_2, x_3) + \cdots + g_{1,2,3,4}(x_1, x_2, x_3, x_4) + \cdots + g_{1,\ldots,p}(x_1, \ldots, x_p)\end{aligned}$$

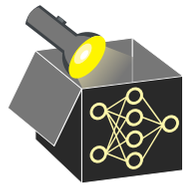
$\rightsquigarrow$ one component for every possible combination S of indices, allowed to formally only depend on these features / be a function of these features

Problems:

- How to find / compute such a decomposition for arbitrary black-box models $\hat{f}$?
- ... such that the decomposition is useful / has nice properties (w.r.t. the model / w.r.t. the data)?

GENERAL FORM OF FUNCTIONAL DECOMPOSITION

RABITZ_2011 | CHASTAING_2012



Definition

Functional decomposition: Additive decomposition of a function $\hat{f} : \mathbb{R}^p \mapsto \mathbb{R}$ into a sum of components of different dimensions w.r.t. inputs $x_1, \ldots, x_p$:

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \sum_{S \subseteq \{1,\ldots,p\}} g_S(\mathbf{x}_S) \\ &= g_\emptyset + g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) + \\ &\quad g_{1,2}(x_1, x_2) + \cdots + g_{p-1,p}(x_{p-1}, x_p) + \cdots + \\ &\quad g_{1,2,3}(x_1, x_2, x_3) + \cdots + g_{1,2,3,4}(x_1, x_2, x_3, x_4) + \cdots + g_{1,\ldots,p}(x_1, \ldots, x_p)\end{aligned}$$

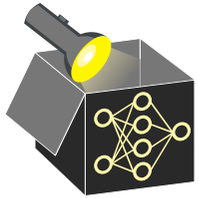
$\rightsquigarrow$ one component for every possible combination S of indices, allowed to formally only depend on these features / be a function of these features

Problems:

- How to find / compute such a decomposition for any black-box models $\hat{f}$?
- ... such that the decomposition is useful / has nice properties (w.r.t. the model / w.r.t. the data)?

GENERAL FORM OF FUNCTIONAL DECOMPOSITION

Li and Rabitz (2011) | Chastaing et al. (2012)



Definition

Functional decomposition: Additive decomposition of a function $\hat{f} : \mathbb{R}^p \mapsto \mathbb{R}$ into a sum of components of different dimensions w.r.t. inputs $x_1, \ldots, x_p$:

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \sum_{S \subseteq \{1,\ldots,p\}} g_S(\mathbf{x}_S) \\ &= g_\emptyset + g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) + \\ &\quad g_{1,2}(x_1, x_2) + \cdots + g_{p-1,p}(x_{p-1}, x_p) + \cdots + \\ &\quad g_{1,2,3}(x_1, x_2, x_3) + \cdots + g_{1,2,3,4}(x_1, x_2, x_3, x_4) + \cdots + g_{1,\ldots,p}(x_1, \ldots, x_p)\end{aligned}$$

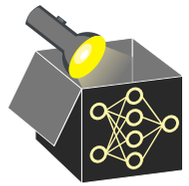
$\rightsquigarrow$ one component for every possible combination S of indices

Sort terms according to degree of interaction:

- $g_\emptyset \hat{=}$ Constant mean (intercept)
- $g_j \hat{=}$ first-order or main effect of j-th feature alone on $\hat{f}(\mathbf{x})$
- $g_{j,k} \hat{=}$ second-order interaction effect of features j and k w.r.t. $\hat{f}(\mathbf{x})$
- $g_S(\mathbf{x}_S) \hat{=}$ $|S|$-order effect, depends **only** on features in S

GENERAL FORM OF FUNCTIONAL DECOMPOSITION

RABITZ_2011 | CHASTAING_2012



Definition

Functional decomposition: Additive decomposition of a function $\hat{f} : \mathbb{R}^p \mapsto \mathbb{R}$ into a sum of components of different dimensions w.r.t. inputs $x_1, \ldots, x_p$:

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \sum_{S \subseteq \{1,\ldots,p\}} g_S(\mathbf{x}_S) \\ &= g_\emptyset + g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) + \\ &\quad g_{1,2}(x_1, x_2) + \cdots + g_{p-1,p}(x_{p-1}, x_p) + \cdots + \\ &\quad g_{1,2,3}(x_1, x_2, x_3) + \cdots + g_{1,2,3,4}(x_1, x_2, x_3, x_4) + \cdots + g_{1,\ldots,p}(x_1, \ldots, x_p)\end{aligned}$$

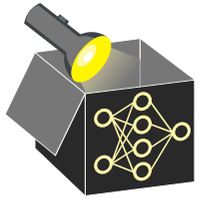
$\rightsquigarrow$ one component for every possible combination S of indices

Sort terms according to degree of interaction:

- $g_\emptyset \hat{=}$ Constant mean (intercept)
- $g_j \hat{=}$ first-order or main effect of j-th feature alone on $\hat{f}(\mathbf{x})$
- $g_{j,k} \hat{=}$ second-order interaction effect of features j and k w.r.t. $\hat{f}(\mathbf{x})$
- $g_S(\mathbf{x}_S) \hat{=}$ $|S|$-order effect, depends **only** on features in S

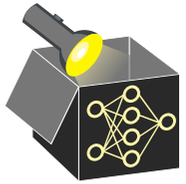
PROPERTIES OF THE DEFINITION

- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure



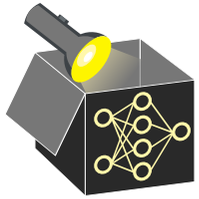
PROPERTIES OF THE DEFINITION

- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure



Interpretable Machine Learning – 7 / 11

PROPERTIES OF THE DEFINITION

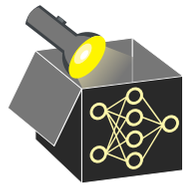


- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled

PROPERTIES OF THE DEFINITION

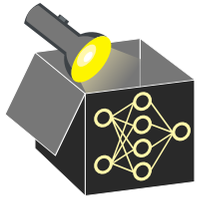


- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled

PROPERTIES OF THE DEFINITION

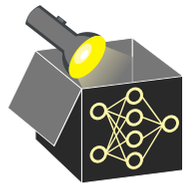


- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled
- Easy decomposition also for decision trees and tree ensembles (see below)

PROPERTIES OF THE DEFINITION

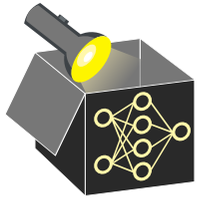


- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled
- Easy decomp. also for decision trees and tree ensembles (see below)

PROPERTIES OF THE DEFINITION

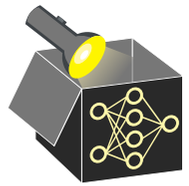


- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled
- Easy decomposition also for decision trees and tree ensembles (see below)
- **Problem 1:** Calculating decomposition extremely difficult, often infeasible
  - For p features: Decomposition with 2^p terms $\rightarrow$ too many different terms, difficult to interpret

PROPERTIES OF THE DEFINITION

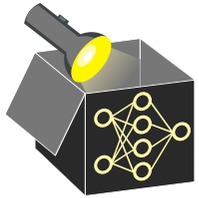


- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled
- Easy decomp. also for decision trees and tree ensembles (see below)
- **Problem 1:** Calculating decomposition extremely difficult, often infeasible
  - For p features: Decomposition with 2^p terms $\rightarrow$ too many different terms, difficult to interpret

PROPERTIES OF THE DEFINITION

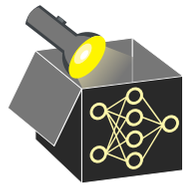


- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled
- Easy decomposition also for decision trees and tree ensembles (see below)
- **Problem 1:** Calculating decomposition extremely difficult, often infeasible
  - For p features: Decomposition with 2^p terms $\rightarrow$ too many different terms, difficult to interpret
- **Problem 2:** Definition not complete: Decomposition not unique, many trivial decompositions not useful
  $\rightarrow$ More requirements or constraints needed to ensure decomposition is meaningful
  - Even worse once features are dependent or correlated (see later)

PROPERTIES OF THE DEFINITION



- **Interpretability**: Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
  $\implies$ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_\emptyset + g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

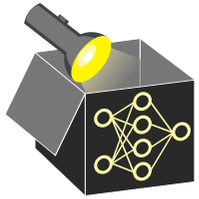
- Same for LMs: Decomposition explicitly modeled
- Easy decomp. also for decision trees and tree ensembles (see below)
- **Problem 1:** Calculating decomposition extremely difficult, often infeasible
  - For p features: Decomposition with 2^p terms $\rightarrow$ too many different terms, difficult to interpret
- **Problem 2:** Definition not complete: Decomposition not unique, many trivial decompositions not useful
  $\rightarrow$ More requirements or constraints needed to ensure decomposition is meaningful
  - Even worse once features are dependent or correlated (see later)

PROBLEM 2: DEFINITION NOT ENOUGH

Example

Again consider

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

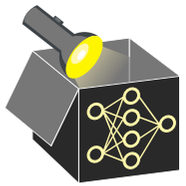


PROBLEM 2: DEFINITION NOT ENOUGH

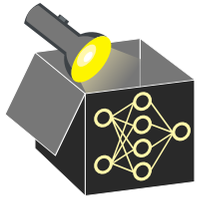
Example

Again consider

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$



PROBLEM 2: DEFINITION NOT ENOUGH



Example

Again consider

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

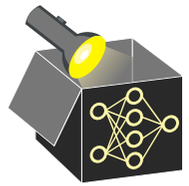
Two possible decompositions (valid according to definition):

$$g_{1,\ldots,p}(x_1, \ldots, x_p) := \hat{f}(\mathbf{x}) \text{ and for all other terms } g_S(\mathbf{x}_S) := 0,$$

or:

$$g_\emptyset = 1\,; \quad g_1(x_1) = x_1\,;\ g_2(x_2) = 2x_2\,;\ g_3(x_3) = 3x_3\,;$$
$$g_{1,2}(x_1, x_2) = \frac{1}{2}x_1x_2\,;\ g_{1,3}(x_1, x_3) = \frac{1}{3}x_1x_3\,;\ g_{2,3}(x_2, x_3) = \frac{2}{3}x_2x_3\,;$$
$$\text{and} \quad g_{1,2,3}(x_1, x_2, x_3) = \hat{f}(x_1, x_2, x_3) - \sum_{S \subsetneq \{1,2,3\}} g_S(\mathbf{x}_S)$$

PROBLEM 2: DEFINITION NOT ENOUGH



Example

Again consider

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

Two possible decompositions (valid according to definition):

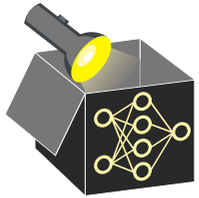
$$g_{1,\ldots,p}(x_1, \ldots, x_p) := \hat{f}(\mathbf{x}) \text{ and for all other terms } g_S(\mathbf{x}_S) := 0,$$

or:

$$g_\emptyset = 1\,; \quad g_1(x_1) = x_1\,;\ g_2(x_2) = 2x_2\,;\ g_3(x_3) = 3x_3\,;$$
$$g_{1,2}(x_1, x_2) = \frac{1}{2}x_1x_2\,;\ g_{1,3}(x_1, x_3) = \frac{1}{3}x_1x_3\,;\ g_{2,3}(x_2, x_3) = \frac{2}{3}x_2x_3\,;$$
$$\text{and} \quad g_{1,2,3}(x_1, x_2, x_3) = \hat{f}(x_1, x_2, x_3) - \sum_{S \subsetneq \{1,2,3\}} g_S(\mathbf{x}_S)$$

Interpretable Machine Learning – 8 / 11

PROBLEM 2: DEFINITION NOT ENOUGH



Example

Again consider

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

Two possible decompositions (valid according to definition):

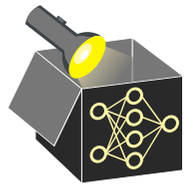
$$g_{1,\ldots,p}(x_1, \ldots, x_p) := \hat{f}(\mathbf{x}) \text{ and for all other terms } g_S(\mathbf{x}_S) := 0,$$

or:

$$g_\emptyset = 1\,; \quad g_1(x_1) = x_1\,;\ g_2(x_2) = 2x_2\,;\ g_3(x_3) = 3x_3\,;$$
$$g_{1,2}(x_1, x_2) = \frac{1}{2}x_1x_2\,;\ g_{1,3}(x_1, x_3) = \frac{1}{3}x_1x_3\,;\ g_{2,3}(x_2, x_3) = \frac{2}{3}x_2x_3\,;$$
$$\text{and} \quad g_{1,2,3}(x_1, x_2, x_3) = \hat{f}(x_1, x_2, x_3) - \sum_{S \subsetneq \{1,2,3\}} g_S(\mathbf{x}_S)$$

$\Longrightarrow$ Definition of decomposition not unique

PROBLEM 2: DEFINITION NOT ENOUGH



Example

Again consider

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

Two possible decompositions (valid according to definition):

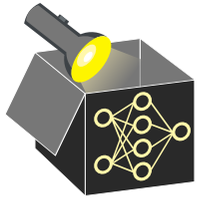
$$g_{1,\ldots,p}(x_1, \ldots, x_p) := \hat{f}(\mathbf{x}) \text{ and for all other terms } g_S(\mathbf{x}_S) := 0,$$

or:

$$g_\emptyset = 1\,; \quad g_1(x_1) = x_1\,;\ g_2(x_2) = 2x_2\,;\ g_3(x_3) = 3x_3\,;$$
$$g_{1,2}(x_1, x_2) = \frac{1}{2}x_1x_2\,;\ g_{1,3}(x_1, x_3) = \frac{1}{3}x_1x_3\,;\ g_{2,3}(x_2, x_3) = \frac{2}{3}x_2x_3\,;$$
$$\text{and} \quad g_{1,2,3}(x_1, x_2, x_3) = \hat{f}(x_1, x_2, x_3) - \sum_{S \subsetneq \{1,2,3\}} g_S(\mathbf{x}_S)$$

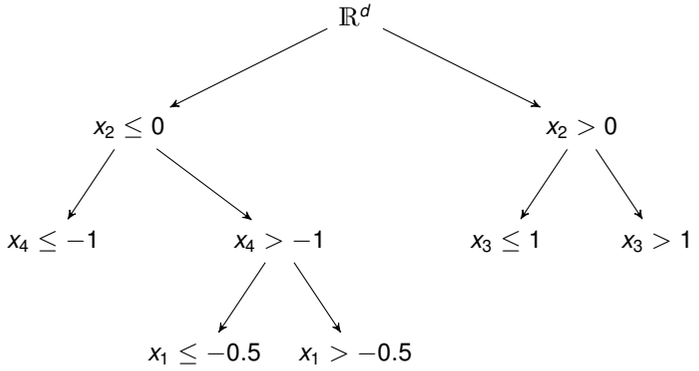
$\Longrightarrow$ Definition of decomposition not unique

EXAMPLE: DECISION TREES

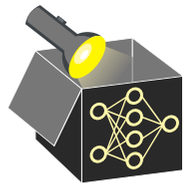


Define *interaction type t* of a node: subset of features involved in constructing this node.

Example:

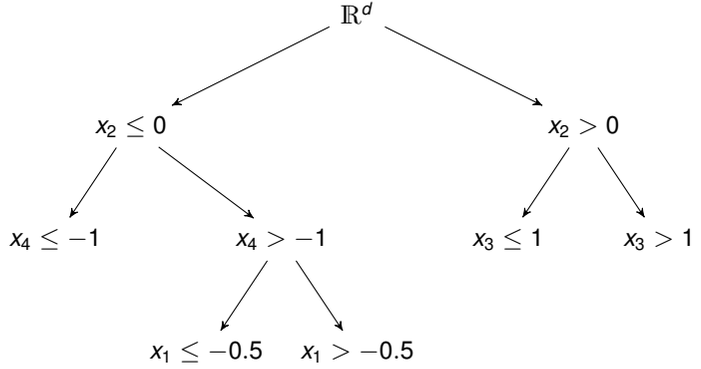


EXAMPLE: DECISION TREES

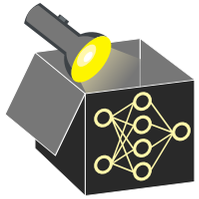


Define *interaction type t* of a node: subset of features used to build this node.

Example:

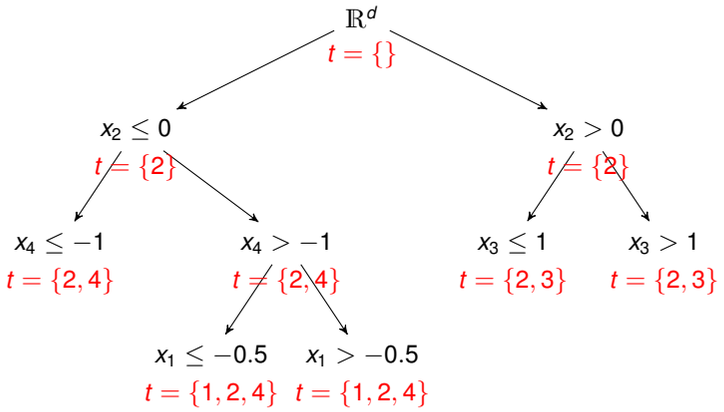


EXAMPLE: DECISION TREES



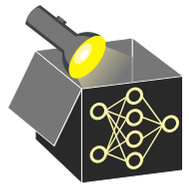
Define *interaction type t* of a node: subset of features involved in constructing this node.

Example:



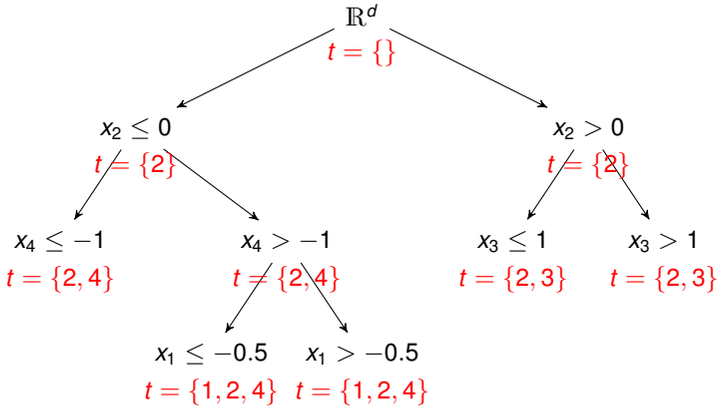
$\Rightarrow$ Degree of interaction in each node is $|t|$.

EXAMPLE: DECISION TREES



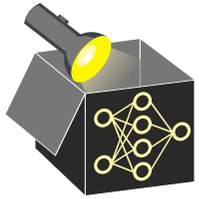
Define *interaction type t* of a node: subset of features used to build this node.

Example:



$\Rightarrow$ Degree of interaction in each node is $|t|$.

DECOMPOSITION FOR DECISION TREES



Here: Decomposition via indicator functions

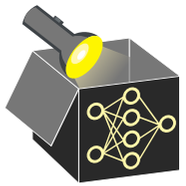
$$\hat{f}(\mathbf{x}) = g_{\emptyset} + g_{2,4}(x_2, x_4) + g_{2,3}(x_2, x_3) + g_{1,2,4}(x_1, x_2, x_4)$$

$\implies$ Decomposition has no main effect, but model certainly contains an effect of e.g. x_2

$\implies$ Lower-order effects "hidden" inside higher-order terms

$\rightsquigarrow$ reading from decision tree not enough, "bad decomposition"

DECOMPOSITION FOR DECISION TREES



Here: Decomposition via indicator functions

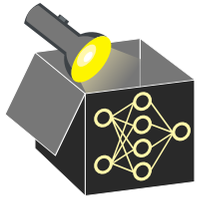
$$\hat{f}(\mathbf{x}) = g_{\emptyset} + g_{2,4}(x_2, x_4) + g_{2,3}(x_2, x_3) + g_{1,2,4}(x_1, x_2, x_4)$$

$\implies$ Decomposition has no main effect, but model certainly contains an effect of e.g. x_2

$\implies$ Lower-order effects "hidden" inside higher-order terms

$\rightsquigarrow$ reading from decision tree not enough, "bad decomposition"

DECOMPOSITION FOR DECISION TREES



Here: Decomposition via indicator functions

$$\hat{f}(\mathbf{x}) = g_{\emptyset} + g_{2,4}(x_2, x_4) + g_{2,3}(x_2, x_3) + g_{1,2,4}(x_1, x_2, x_4)$$

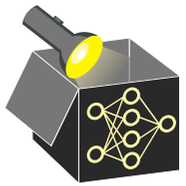
$\implies$ Decomposition has no main effect, but model certainly contains an effect of e.g. x_2

$\implies$ Lower-order effects "hidden" inside higher-order terms

$\rightsquigarrow$ reading from decision tree not enough, "bad decomposition"

Note: ▸ Yang (2024) propose a (quite complicated) solution for this case

DECOMPOSITION FOR DECISION TREES



Here: Decomposition via indicator functions

$$\hat{f}(\mathbf{x}) = g_{\emptyset} + g_{2,4}(x_2, x_4) + g_{2,3}(x_2, x_3) + g_{1,2,4}(x_1, x_2, x_4)$$

$\implies$ Decomposition has no main effect, but model certainly contains an effect of e.g. x_2

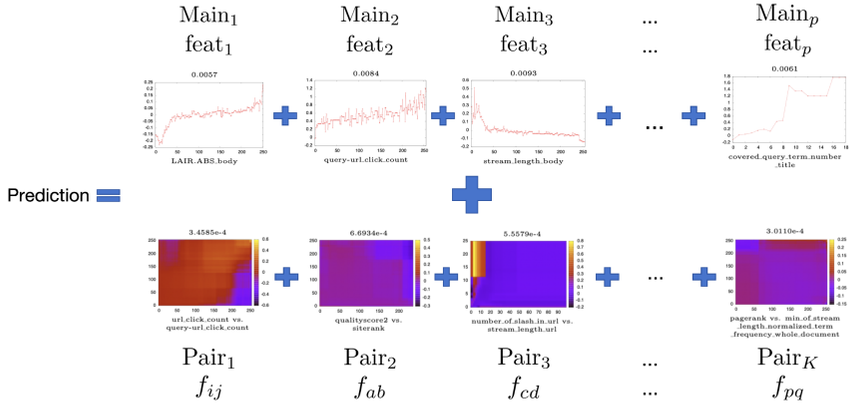
$\implies$ Lower-order effects "hidden" inside higher-order terms

$\rightsquigarrow$ reading from decision tree not enough, "bad decomposition"

Note: ▸ Yang 2024 propose a (quite complicated) solution for this case

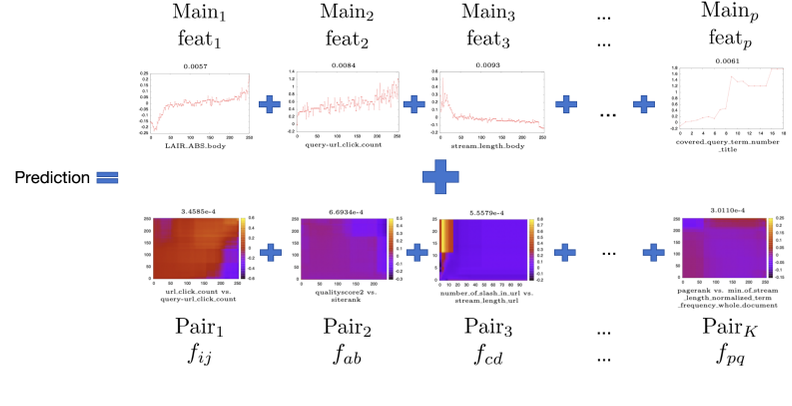
EXAMPLE: EBM

- See before: **GAMs** have functional decomposition by definition
- **EBMs:** Sum of the final one- and two-dimensional components

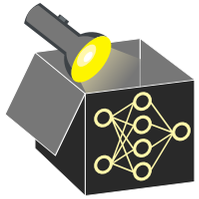


EXAMPLE: EBM

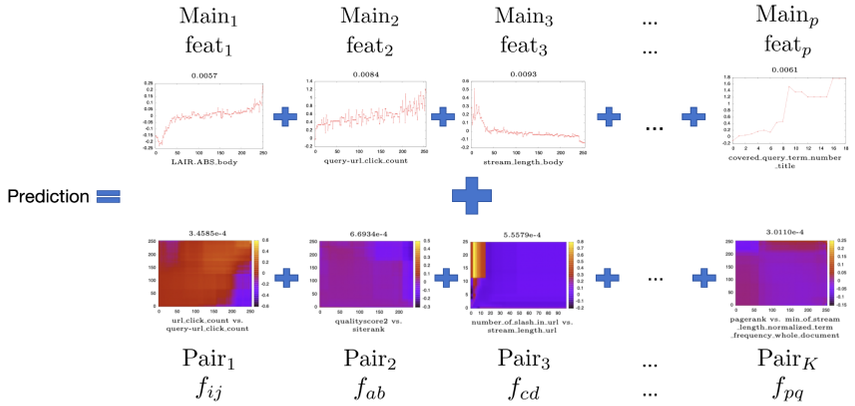
- See before: **GAMs** have functional decomposition by definition
- **EBMs:** Sum of the final one- and two-dimensional components



EXAMPLE: EBM

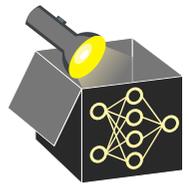


- See before: **GAMs** have functional decomposition by definition
- **EBMs:** Sum of the final one- and two-dimensional components

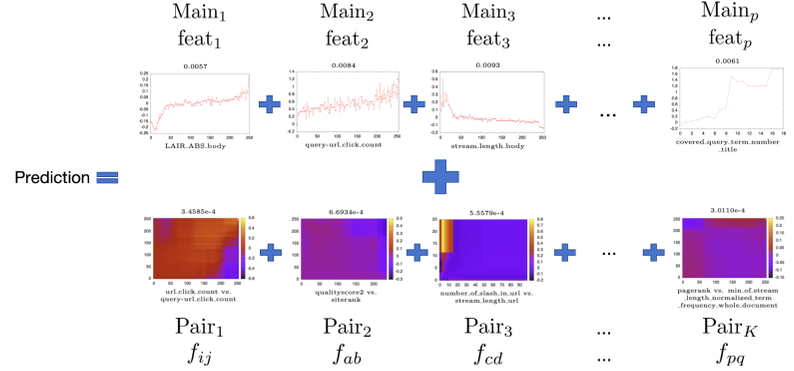


- In general: Model with functional decomposition up to max. order 2 is always “inherently interpretable”
- **Reason:** Visualization of all components

EXAMPLE: EBM



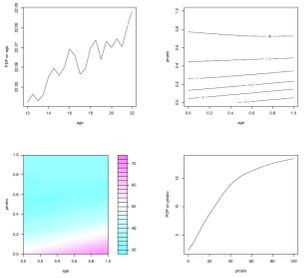
- See before: **GAMs** have functional decomposition by definition
- **EBMs:** Sum of the final one- and two-dimensional components



- In general: Model with functional decomposition up to max. order 2 is always “inherently interpretable”
- **Reason:** Visualization of all components

Interpretable Machine Learning

Functional ANOVA

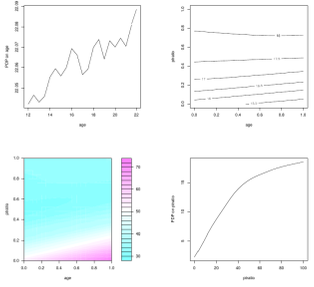


Learning goals

- One method for functional decomposition: Classical functional ANOVA (fANOVA)
- Algorithm for calculating the components in a fANOVA
- Variance decomposition in fANOVA

Interpretable Machine Learning

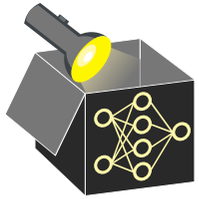
Functional Decompositions
Functional ANOVA



Learning goals

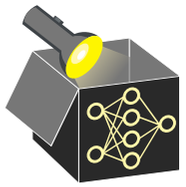
- One method for functional decomposition: Classical functional ANOVA (fANOVA)
- Algorithm for calculating the components in a fANOVA
- Variance decomposition in fANOVA

INTRODUCTION AND HISTORY OF FANOVA



- One possible method to obtain functional decomposition
- Since 1940's: Developed under different names in mathematics and sensitivity analysis
- Since 1990's: Developed for probability distributions or statistical data
- Since 2000's: Applied to machine learning, subsequently alternatives developed extending applicability
- **Assumption**: Independent features

INTRODUCTION AND HISTORY OF FANOVA

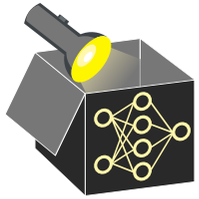


- One possible method to obtain functional decomposition
- Since 1940's: Developed under different names in mathematics and sensitivity analysis
- Since 1990's: Developed for probability distributions or statistical data
- Since 2000's: Applied to machine learning, subsequently alternatives developed extending applicability
- **Assumption**: Independent features

STANDARD FANOVA: IDEA

- Example:

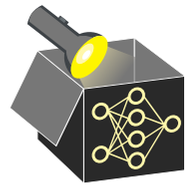
$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$



STANDARD FANOVA: IDEA

- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

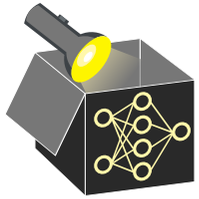


STANDARD FANOVA: IDEA

- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.

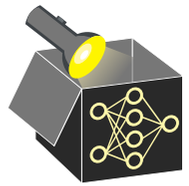


STANDARD FANOVA: IDEA

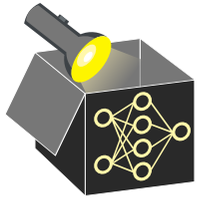
- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.



STANDARD FANOVA: IDEA

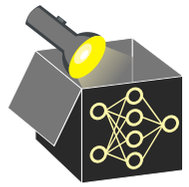


- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In first step, compute main effects using feature effect methods
  Here: PDP + more general PD-functions

STANDARD FANOVA: IDEA

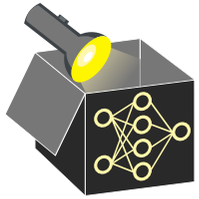


- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In 1st step, compute main effects using feat effect methods
  Here: PDP + more general PD-functions

STANDARD FANOVA: IDEA



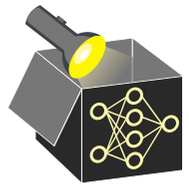
- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In first step, compute main effects using feature effect methods
  Here: PDP + more general PD-functions
- **Idea for fANOVA:** PD-function $\hat{f}_{S;PD}$ = sum of all components $g_{\tilde{S}}$ up to this order

$$\hat{f}_{S;PD}(\mathbf{x}_S) \;=\; \sum_{V \subseteq S} g_V(\mathbf{x}_V)$$

STANDARD FANOVA: IDEA



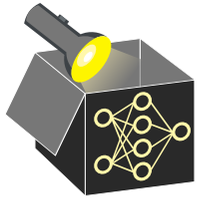
- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In 1st step, compute main effects using feat effect methods
  Here: PDP + more general PD-functions
- **Idea for fANOVA:** PD-function $\hat{f}_{S;PD}$ = sum of all components $g_{\tilde{S}}$ up to this order

$$\hat{f}_{S;PD}(\mathbf{x}_S) \;=\; \sum_{V \subseteq S} g_V(\mathbf{x}_V)$$

STANDARD FANOVA: IDEA



- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In first step, compute main effects using feature effect methods
  Here: PDP + more general PD-functions
- **Idea for fANOVA:** PD-function $\hat{f}_{S;PD}$ = sum of all components $g_{\tilde{S}}$ up to this order

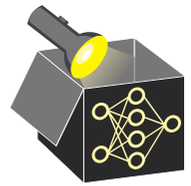
$$\hat{f}_{S;PD}(\mathbf{x}_S) \;=\; \sum_{V \subseteq S} g_V(\mathbf{x}_V)$$

- **Remember:**
  Idea of PDPs or general PD-functions: Average out all other features

$\Rightarrow$ Total formula for calculating the components g_S in the fANOVA algorithm:

$$g_S(\mathbf{x}_S) \;=\; (\text{average out all features not contained in } S) \\ -\; (\text{All lower-order components})$$

STANDARD FANOVA: IDEA



- Example:

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In 1st step, compute main effects using feat effect methods
  Here: PDP + more general PD-functions
- **Idea for fANOVA:** PD-function $\hat{f}_{S;PD}$ = sum of all components $g_{\tilde{S}}$ up to this order

$$\hat{f}_{S;PD}(\mathbf{x}_S) \;=\; \sum_{V \subseteq S} g_V(\mathbf{x}_V)$$

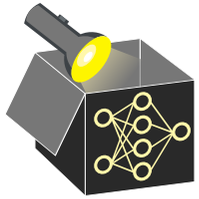
- **Remember:**
  Idea of PDPs or general PD-functions: Average out all other features

$\Rightarrow$ Total formula for calculating the components g_S in the fANOVA algorithm:

$$g_S(\mathbf{x}_S) \;=\; (\text{average out all features not contained in } S) \\ -\; (\text{All lower-order components})$$

FORMAL DEFINITION AND COMPUTATION

Hooker (2004)



Definition

Recursive computation using PD-functions
(here $-S = \{1, \dots, p\} \setminus S$ denotes all indices not contained in S):

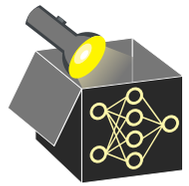
$$g_S(\mathbf{x}_S) = \hat{f}_{S;PD}(\mathbf{x}_S) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \mathbb{E}_{\mathbf{x}_{-S}} \left[\hat{f}(\mathbf{x}_S, \mathbf{X}_{-S}) \right] - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

$$= \int \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}) d\mathbb{P}(\mathbf{x}_{-S}) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

- Expectation integrates $\hat{f}(\mathbf{x})$ over all input features except $\mathbf{x}_S$
- Subtract sum of g_V to remove all lower-order effects and center the effect
- **Note:** If no distribution given: Uniform distribution or plain integral

FORMAL DEFINITION AND COMPUTATION

HOOKER_2004



Definition

Recursive computation using PD-functions
(here $-S = \{1, \dots, p\} \setminus S$ denotes all indices not contained in S):

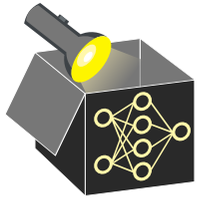
$$g_S(\mathbf{x}_S) = \hat{f}_{S;PD}(\mathbf{x}_S) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \mathbb{E}_{-S} \left[\hat{f}(\mathbf{x}_{S,-S}) \right] - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

$$= \int \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}) d\mathbb{P}(\mathbf{x}_{-S}) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

- Expectation integrates $\hat{f}(\mathbf{x})$ over all input features except $\mathbf{x}_S$
- Subtract sum of g_V to remove all lower-order effects and center the effect
- **Note:** If no distribution given: Uniform distribution or plain integral

FORMAL DEFINITION AND COMPUTATION

Hooker (2004)



Definition

Recursive computation using PD-functions
(here $-S = \{1, \ldots, p\} \setminus S$ denotes all indices not contained in S):

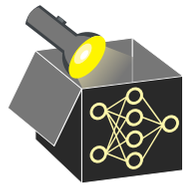
$$g_S(\mathbf{x}_S) = \hat{f}_{S;PD}(\mathbf{x}_S) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \mathbb{E}_{\mathbf{X}_{-S}}\left[\hat{f}(\mathbf{x}_S, \mathbf{X}_{-S})\right] - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

- Recursive computation:

$$\begin{aligned}
g_\emptyset &= \mathbb{E}_{\mathbf{X}}\left[\hat{f}(\mathbf{X})\right] \\
g_j(x_j) &= \mathbb{E}_{\mathbf{X}_{-j}}\left[\hat{f}(\mathbf{X}) \mid X_j = x_j\right] - g_\emptyset, \ \forall j \in \{1, \ldots, p\} \\
&\vdots \\
g_{1,\ldots,p}(\mathbf{x}) &= \hat{f}(\mathbf{x}) - \textstyle\sum_{S \subsetneq \{1,\ldots,p\}} g_S(\mathbf{x}_S) \\
&= \hat{f}(\mathbf{x}) - g_{1,\ldots,p-1}(x_1, \ldots x_{p-1}) - \cdots - g_{1,2}(x_1, x_2) \\
&\quad - g_p(x_p) - \cdots - g_2(x_2) - g_1(x_1) - g_\emptyset
\end{aligned}$$

FORMAL DEFINITION AND COMPUTATION

HOOKER_2004



Definition

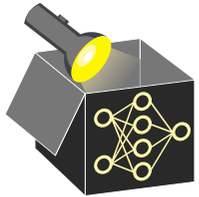
Recursive computation using PD-functions
(here $-S = \{1, \ldots, p\} \setminus S$ denotes all indices not contained in S):

$$g_S(\mathbf{x}_S) = \hat{f}_{S;PD}(\mathbf{x}_S) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \mathbb{E}_{-S}\left[\hat{f}(\mathbf{x}_S, -S)\right] - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

- Recursive computation:

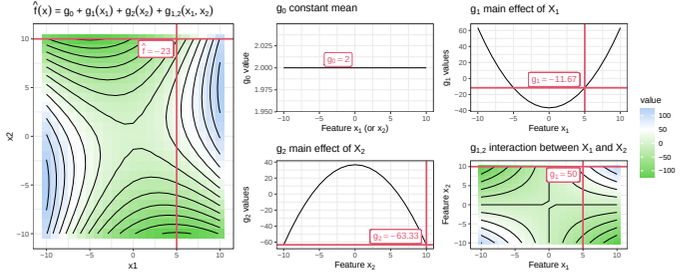
$$\begin{aligned}
g_\emptyset &= \mathbb{E}_{[\hat{f}()]} \\
g_j(x_j) &= \mathbb{E}_{-j}\left[\hat{f}() \mid X_j = x_j\right] - g_\emptyset, \ \forall j \in \{1, \ldots, p\} \\
&\vdots \\
g_{1,\ldots,p}(\mathbf{x}) &= \hat{f}(\mathbf{x}) - \textstyle\sum_{S \subsetneq \{1,\ldots,p\}} g_S(\mathbf{x}_S) \\
&= \hat{f}(\mathbf{x}) - g_{1,\ldots,p-1}(x_1, \ldots x_{p-1}) - \cdots - g_{1,2}(x_1, x_2) \\
&\quad - g_p(x_p) - \cdots - g_2(x_2) - g_1(x_1) - g_\emptyset
\end{aligned}$$

STANDARD FANOVA – EXAMPLE



Example: $\hat{f}(\mathbf{x}) = 2 + x_1^2 - x_2^2 + x_1 \cdot x_2$ (e.g., for $x_1 = 5$ and $x_2 = 10$ we have $\hat{f}(\mathbf{x}) = -23$)

- Computation of components using feature values $x_1 = x_2 = (-10, -9, \ldots, 10)^\top$ gives:

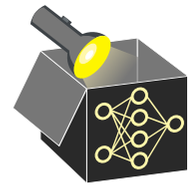


For $x_1 = 5$ and $x_2 = 10$:

- $g_\emptyset = 2$
- $g_1(x_1) = -9.67$
- $g_2(x_2) = -65.33$
- $g_{1,2}(x_1, x_2) = 50$

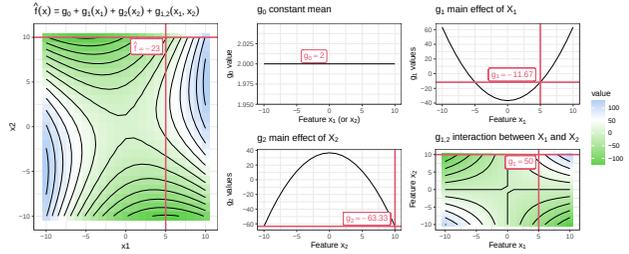
$\Rightarrow \hat{f}(\mathbf{x}) = -23$

STANDARD FANOVA – EXAMPLE



Example: $\hat{f}(\mathbf{x}) = 2 + x_1^2 - x_2^2 + x_1 \cdot x_2$ (e.g., for $x_1 = 5$ and $x_2 = 10$ we have $\hat{f}(\mathbf{x}) = -23$)

- Computation of components using feature values $x_1 = x_2 = (-10, -9, \ldots, 10)^\top$ gives:



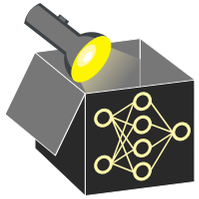
For $x_1 = 5$ and $x_2 = 10$:

- $g_\emptyset = 2$
- $g_1(x_1) = -9.67$
- $g_2(x_2) = -65.33$
- $g_{1,2}(x_1, x_2) = 50$

$\Rightarrow \hat{f}(\mathbf{x}) = -23$

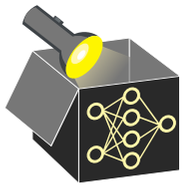
STANDARD FANOVA - EXAMPLE

In-class task

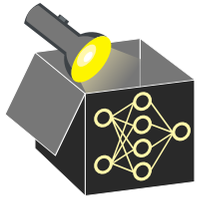


STANDARD FANOVA - EXAMPLE

In-class task



STANDARD FANOVA - EXAMPLE REVISITED



Example

$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad$ uniformly distributed

- **Intercept:**

$$\begin{aligned} g_\emptyset &= \mathbb{E}\left[\hat{f}(x_1, x_2)\right] = \int_0^1 \int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_1 \, dx_2 \\ &= 4 - \left(\int_0^1 2x_1 \, dx_1\right) + \left(\int_0^1 0.3e^{x_2} \, dx_2\right) + \left(\int_0^1 |x_1| \, dx_1\right)\left(\int_0^1 x_2 \, dx_2\right) \\ &= 4 - 1 + 0.3(e-1) + 0.5^2 = 2.95 + 0.3e. \end{aligned}$$

STANDARD FANOVA - EXAMPLE REVISITED

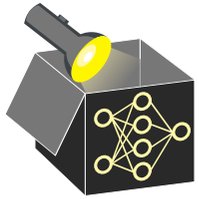
Example

$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad$ uniformly distributed

- **Intercept:**

$$\begin{aligned} g_\emptyset &= \mathbb{E}\left[\hat{f}(x_1, x_2)\right] = \int_0^1 \int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_1 \, dx_2 \\ &= 4 - \left(\int_0^1 2x_1 \, dx_1\right) + \left(\int_0^1 0.3e^{x_2} \, dx_2\right) + \left(\int_0^1 |x_1| \, dx_1\right)\left(\int_0^1 x_2 \, dx_2\right) \\ &= 4 - 1 + 0.3(e-1) + 0.5^2 = 2.95 + 0.3e. \end{aligned}$$

STANDARD FANOVA - EXAMPLE REVISITED



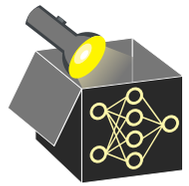
Example

$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad$ uniformly distributed

- **First-order components:**

$$
\begin{aligned}
g_1(x_1) = \hat{f}_{1;PD}(x_1) - g_\emptyset &= \left(\int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_2\right) - g_\emptyset \\
&= 4 - 2x_1 + 0.3(e-1) + |x_1| \cdot \frac{1}{2} - (2.95 + 0.3e) \\
&= -2x_1 + 0.5|x_1| + 0.75 \\
g_2(x_2) = \hat{f}_{2;PD}(x_2) - g_\emptyset &= \left(\int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_1\right) - g_\emptyset \\
&= 4 - 1 + 0.3e^{x_2} + \frac{1}{2} \cdot x_2 - (2.95 + 0.3e) \\
&= 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05
\end{aligned}
$$

STANDARD FANOVA - EXAMPLE REVISITED



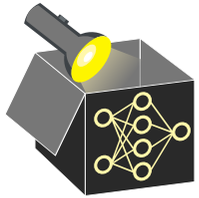
Example

$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad$ uniformly distributed

- **First-order components:**

$$
\begin{aligned}
g_1(x_1) = \hat{f}_{1;PD}(x_1) - g_\emptyset &= \left(\int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_2\right) - g_\emptyset \\
&= 4 - 2x_1 + 0.3(e-1) + |x_1| \cdot \frac{1}{2} - (2.95 + 0.3e) \\
&= -2x_1 + 0.5|x_1| + 0.75 \\
g_2(x_2) = \hat{f}_{2;PD}(x_2) - g_\emptyset &= \left(\int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_1\right) - g_\emptyset \\
&= 4 - 1 + 0.3e^{x_2} + \frac{1}{2} \cdot x_2 - (2.95 + 0.3e) \\
&= 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05
\end{aligned}
$$

STANDARD FANOVA - EXAMPLE REVISITED



Example

$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad$ uniformly distributed

- **Second-order component:**

$$\begin{aligned} g_{12}(x_1, x_2) &= \hat{f}_{\{1,2\};PD}(x_1, x_2) - g_\emptyset - g_1(x_1) - g_2(x_2) \\ &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - (2.95 + 0.3e) \\ &\quad - (-2x_1 + 0.5|x_1| + 0.75) - (0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05) \\ &= |x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25 \end{aligned}$$

$\Rightarrow$ All components shifted to have mean 0

$\Rightarrow$ Parts of $|x_1|x_2$, which intuitively seems to be the "interaction term", is attributed to the main effects (correctly, depends on distribution!)

STANDARD FANOVA - EXAMPLE REVISITED

Example

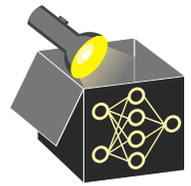
$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad$ uniformly distributed

- **Second-order component:**

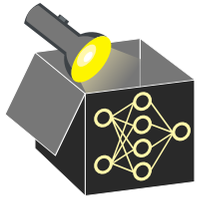
$$\begin{aligned} g_{12}(x_1, x_2) &= \hat{f}_{\{1,2\};PD}(x_1, x_2) - g_\emptyset - g_1(x_1) - g_2(x_2) \\ &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - (2.95 + 0.3e) \\ &\quad - (-2x_1 + 0.5|x_1| + 0.75) - (0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05) \\ &= |x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25 \end{aligned}$$

$\Rightarrow$ All components shifted to have mean 0

$\Rightarrow$ Parts of $|x_1|x_2$, which intuitively seems to be the "interaction term", is attributed to the main effects (correctly, depends on distribution!)



ESTIMATE FANOVA IN PRACTICE



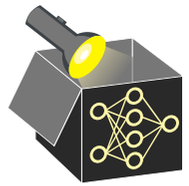
Main part: Calculate all PD-functions $\rightarrow 2^p$ **many PD-functions**
Estimation of a single PD-function: **Sampling**
(so-called **Monte-Carlo integration**)

- Same idea as for PDPs: Fix **grid values** for features x_S
  Here: Same grid for all features over the whole algorithm
- Estimate integral by sampling: for grid value $\mathbf{x}_S^*$:

$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \mathbb{E}_{\mathbf{X}_{-S}}\left[\hat{f}(\mathbf{x}_S^*, \mathbf{X}_{-S})\right] \approx \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$$

- Or: for each grid value $\mathbf{x}_S^*$, sample only $n_s < n$ many random samples (e.g. sampling uniformly)

ESTIMATE FANOVA IN PRACTICE



Main part: Calculate all PD-functions $\rightarrow 2^p$ **many PD-functions**
Estimation of a single PD-function: **Sampling**
(so-called **Monte-Carlo integration**)

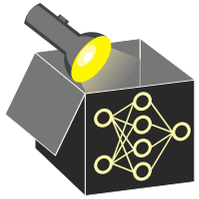
- Same idea as for PDPs: Fix **grid values** for features x_S
  Here: Same grid for all features over the whole algorithm
- Estimate integral by sampling: for grid value $\mathbf{x}_S^*$:

$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \mathbb{E}_{-S}\left[\hat{f}(\mathbf{x}_{S,-S}^*)\right] \approx \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$$

- Or: for each grid value $\mathbf{x}_S^*$, sample only $n_s < n$ many random samples (e.g. sampling uniformly)

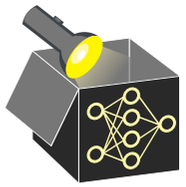
VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?

- Decomposition of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)

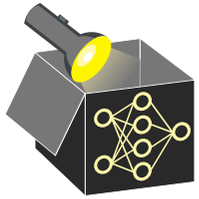


VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?

- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)



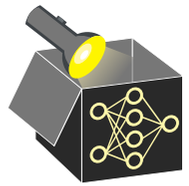
VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomposition of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

$$\begin{aligned}\text{Var}\left[\hat{f}(\mathbf{x})\right] &= \text{Var}\left[g_\emptyset + g_1(x_1) + \ \ldots\ + g_{1,2}(x_1, x_2) + \ \ldots\ + g_{1,\ldots,p}(\mathbf{x})\right] \\ &= \text{Var}\left[g_\emptyset\right] + \text{Var}\left[g_1(x_1)\right] + \ \ldots\ + \text{Var}\left[g_{1,2}(x_1, x_2)\right] + \ \ldots\ + \text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]\end{aligned}$$

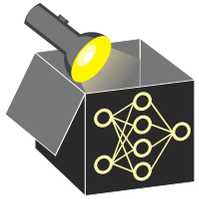
VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

$$\begin{aligned}\text{Var}\left[\hat{f}(\mathbf{x})\right] &= \text{Var}\left[g_\emptyset + g_1(x_1) + \ \ldots\ + g_{1,2}(x_1, x_2) + \ \ldots\ + g_{1,\ldots,p}(\mathbf{x})\right] \\ &= \text{Var}\left[g_\emptyset\right] + \text{Var}\left[g_1(x_1)\right] + \ \ldots\ + \text{Var}\left[g_{1,2}(x_1, x_2)\right] + \ \ldots\ + \text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]\end{aligned}$$

VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



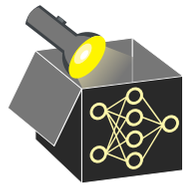
- Decomposition of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

$$\begin{aligned}\text{Var}\left[\hat{f}(\mathbf{x})\right] &= \text{Var}\left[g_\emptyset + g_1(x_1) + \ \ldots\ + g_{1,2}(x_1, x_2) + \ \ldots\ + g_{1,\ldots,p}(\mathbf{x})\right] \\ &= \text{Var}\left[g_\emptyset\right] + \text{Var}\left[g_1(x_1)\right] + \ \ldots\ + \text{Var}\left[g_{1,2}(x_1, x_2)\right] + \ \ldots\ + \text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]\end{aligned}$$

- In other words: Single components uncorrelated (see later)

$$1 = \frac{\text{Var}\left[g_\emptyset\right]}{Var\left[\hat{f}(\mathbf{x})\right]} + \frac{\text{Var}\left[g_1(x_1)\right]}{Var\left[\hat{f}(\mathbf{x})\right]} + \ \ldots\ \frac{\text{Var}\left[g_{1,2}(x_1, x_2)\right]}{Var\left[\hat{f}(\mathbf{x})\right]}\ \ldots\ + \frac{\text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]}{Var\left[\hat{f}(\mathbf{x})\right]}$$

VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

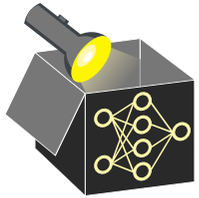
$$\begin{aligned}\text{Var}\left[\hat{f}(\mathbf{x})\right] &= \text{Var}\left[g_\emptyset + g_1(x_1) + \ \ldots\ + g_{1,2}(x_1, x_2) + \ \ldots\ + g_{1,\ldots,p}(\mathbf{x})\right] \\ &= \text{Var}\left[g_\emptyset\right] + \text{Var}\left[g_1(x_1)\right] + \ \ldots\ + \text{Var}\left[g_{1,2}(x_1, x_2)\right] + \ \ldots\ + \text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]\end{aligned}$$

- In other words: Single components uncorrelated (see later)

$$1 = \frac{\text{Var}\left[g_\emptyset\right]}{Var\left[\hat{f}(\mathbf{x})\right]} + \frac{\text{Var}\left[g_1(x_1)\right]}{Var\left[\hat{f}(\mathbf{x})\right]} + \ \ldots\ \frac{\text{Var}\left[g_{1,2}(x_1, x_2)\right]}{Var\left[\hat{f}(\mathbf{x})\right]}\ \ldots\ + \frac{\text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]}{Var\left[\hat{f}(\mathbf{x})\right]}$$

Interpretable Machine Learning – 9 / 9

VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomposition of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

$$\begin{aligned}\text{Var}\left[\hat{f}(\mathbf{x})\right] &= \text{Var}\left[g_\emptyset + g_1(x_1) + \ \ldots\ + g_{1,2}(x_1, x_2) + \ \ldots\ + g_{1,\ldots,p}(\mathbf{x})\right] \\ &= \text{Var}\left[g_\emptyset\right] + \text{Var}\left[g_1(x_1)\right] + \ \ldots\ + \text{Var}\left[g_{1,2}(x_1, x_2)\right] + \ \ldots\ + \text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]\end{aligned}$$

- In other words: Single components uncorrelated (see later)

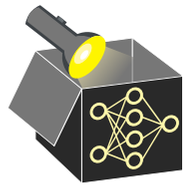
$$1 = \frac{\text{Var}\left[g_\emptyset\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]} + \frac{\text{Var}\left[g_1(x_1)\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]} + \ \ldots\ \frac{\text{Var}\left[g_{1,2}(x_1, x_2)\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]}\ \ldots\ + \frac{\text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]}$$

$\rightarrow$ **Sobol index**: Fraction of variance explained by some component $g_V(\mathbf{x}_V)$:

$$S_V = \frac{\text{Var}\left[g_V(\mathbf{x}_V)\right]}{\text{Var}\left[\hat{f}(\mathbf{x})\right]}$$

$\rightsquigarrow$ Usable as importance measure of component $g_V(\mathbf{x}_V)$

VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

$$\begin{aligned}\text{Var}\left[\hat{f}(\mathbf{x})\right] &= \text{Var}\left[g_\emptyset + g_1(x_1) + \ \ldots\ + g_{1,2}(x_1, x_2) + \ \ldots\ + g_{1,\ldots,p}(\mathbf{x})\right] \\ &= \text{Var}\left[g_\emptyset\right] + \text{Var}\left[g_1(x_1)\right] + \ \ldots\ + \text{Var}\left[g_{1,2}(x_1, x_2)\right] + \ \ldots\ + \text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]\end{aligned}$$

- In other words: Single components uncorrelated (see later)

$$1 = \frac{\text{Var}\left[g_\emptyset\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]} + \frac{\text{Var}\left[g_1(x_1)\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]} + \ \ldots\ \frac{\text{Var}\left[g_{1,2}(x_1, x_2)\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]}\ \ldots\ + \frac{\text{Var}\left[g_{1,\ldots,p}(\mathbf{x})\right]}{\mathit{Var}\left[\hat{f}(\mathbf{x})\right]}$$

$\rightarrow$ **Sobol index**: Fraction of variance explained by some component $g_V(\mathbf{x}_V)$:

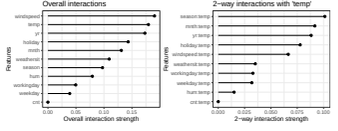
$$S_V = \frac{\text{Var}\left[g_V(\mathbf{x}_V)\right]}{\text{Var}\left[\hat{f}(\mathbf{x})\right]}$$

$\rightsquigarrow$ Usable as importance measure of component $g_V(\mathbf{x}_V)$

Interpretable Machine Learning – 9 / 9

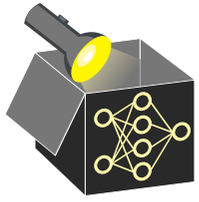
Interpretable Machine Learning

Friedman's H-Statistic



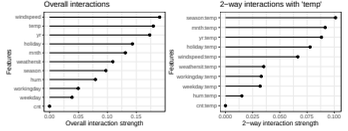
Learning goals

- Friedman's H-statistic with two purposes:
- Measure general *k*-way interactions between arbitrary features
- Measure a single feature's overall interaction strength



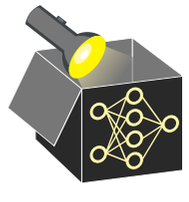
Interpretable Machine Learning

Functional Decompositions
Friedman's H-Statistic



Learning goals

- Friedman's H-statistic with two purposes:
- Measure general *k*-way interactions between arbitrary features
- Measure a single feature's overall interaction strength

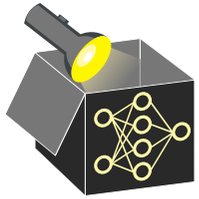


IDEA

Friedman and Popescu (2008)

2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0

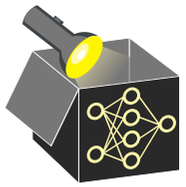


IDEA

POPESCU_2008

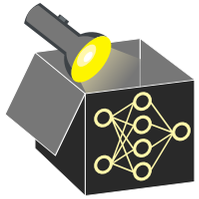
2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0



IDEA

Friedman and Popescu (2008)



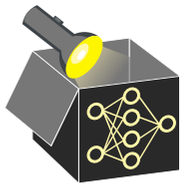
2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

IDEA

POPESCU_2008



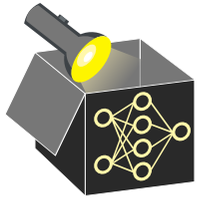
2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

IDEA

Friedman and Popescu (2008)



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

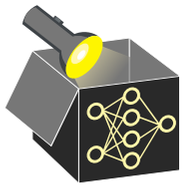
$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

IDEA

POPESCU_2008



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

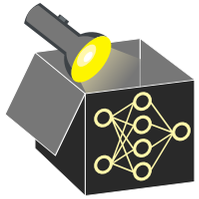
$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

IDEA

Friedman and Popescu (2008)



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

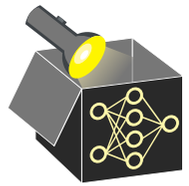
$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- **Definition:** A function $\hat{f}$ contains no 2-way interactions between j and k, if there exists a decomposition

$$\begin{aligned} & \hat{f}^c_{\{jk\},PD}(x_x j, x_k) = g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \quad & \hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \end{aligned}$$

IDEA

POPESCU_2008



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

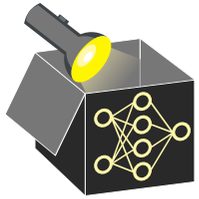
$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- **Definition:** A function $\hat{f}$ contains no 2-way interactions between j and k, if there exists a decomposition

$$\begin{aligned} & \hat{f}^c_{\{jk\},PD}(x_x j, x_k) = g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \quad & \hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \end{aligned}$$

IDEA

Friedman and Popescu (2008)



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

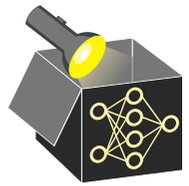
- **Definition:** A function $\hat{f}$ contains no 2-way interactions between j and k, if there exists a decomposition

$$\begin{aligned} & \hat{f}^c_{\{jk\},PD}(x_x j, x_k) = g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \quad & \hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \end{aligned}$$

- This means: There are interactions
  $\Leftrightarrow$ Every possible decomposition must contain some non-zero term $g_{\{j,k\}}(x_j, x_k)$

IDEA

POPESCU_2008



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

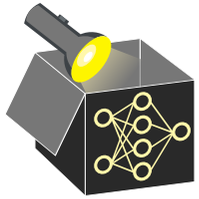
- **Definition:** A function $\hat{f}$ contains no 2-way interactions between j and k, if there exists a decomposition

$$\begin{aligned} & \hat{f}^c_{\{jk\},PD}(x_x j, x_k) = g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \quad & \hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \end{aligned}$$

- This means: There are interactions
  $\Leftrightarrow$ Every possible decomp. must contain some non 0 term $g_{\{j,k\}}(x_j, x_k)$

IDEA

Friedman and Popescu (2008)



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

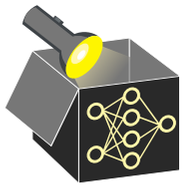
- **Definition:** A function $\hat{f}$ contains no 2-way interactions between j and k, if there exists a decomposition

$$\begin{aligned} & \hat{f}^c_{\{jk\},PD}(x_x j, x_k) = g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \quad & \hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \end{aligned}$$

- This means: There are interactions
  $\Leftrightarrow$ Every possible decomposition must contain some non-zero term $g_{\{j,k\}}(x_j, x_k)$
- Again: remember GAMs

IDEA

POPESCU_2008



2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_\emptyset + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_\emptyset$

$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

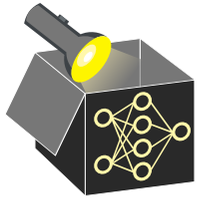
- **Definition:** A function $\hat{f}$ contains no 2-way interactions between j and k, if there exists a decomposition

$$\begin{aligned} & \hat{f}^c_{\{jk\},PD}(x_x j, x_k) = g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \quad & \hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \end{aligned}$$

- This means: There are interactions
  $\Leftrightarrow$ Every possible decomp. must contain some non 0 term $g_{\{j,k\}}(x_j, x_k)$
- Again: remember GAMs

IDEA

Friedman and Popescu (2008)



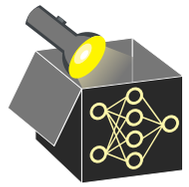
3-way interaction:

- **Definition:** $\hat{f}$ contains no 3-way interactions between features i, j, k, if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

$$\hat{f}_{\{ijk\},PD}(x_i, x_j, x_k) = g_\emptyset + g_i(x_i) + g_j(x_j) + g_k(x_k) + g_{\{i,j\}}(x_i, x_j) + g_{\{i,k\}}(x_i, x_k) + g_{\{i,j\}}(x_j, x_k)$$

IDEA

POPESCU_2008



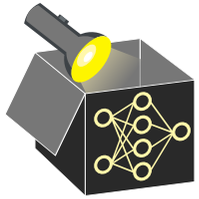
3-way interaction:

- **Definition:** $\hat{f}$ contains no 3-way interactions between features i, j, k, if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

$$\hat{f}_{\{ijk\},PD}(x_i, x_j, x_k) = g_\emptyset + g_i(x_i) + g_j(x_j) + g_k(x_k) + g_{\{i,j\}}(x_i, x_j) + g_{\{i,k\}}(x_i, x_k) + g_{\{i,j\}}(x_j, x_k)$$

IDEA

Friedman and Popescu (2008)



3-way interaction:

- **Definition:** $\hat{f}$ contains no 3-way interactions between features i, j, k, if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

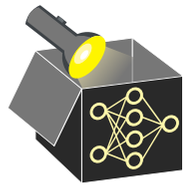
$$\hat{f}_{\{ijk\},PD}(x_i, x_j, x_k) = g_\emptyset + g_i(x_i) + g_j(x_j) + g_k(x_k) + g_{\{i,j\}}(x_i, x_j) + g_{\{i,k\}}(x_i, x_k) + g_{\{i,j\}}(x_j, x_k)$$

- **Example**:

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

IDEA

POPESCU_2008



3-way interaction:

- **Definition:** $\hat{f}$ contains no 3-way interactions between features i, j, k, if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

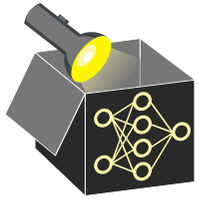
$$\hat{f}_{\{ijk\},PD}(x_i, x_j, x_k) = g_\emptyset + g_i(x_i) + g_j(x_j) + g_k(x_k) + g_{\{i,j\}}(x_i, x_j) + g_{\{i,k\}}(x_i, x_k) + g_{\{i,j\}}(x_j, x_k)$$

- **Example**:

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

IDEA

Friedman and Popescu (2008)



3-way interaction:

- **Definition:** $\hat{f}$ contains no 3-way interactions between features i, j, k, if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

$$\begin{aligned}\hat{f}_{\{ijk\},PD}(x_i, x_j, x_k) = g_\emptyset + g_i(x_i) + g_j(x_j) + g_k(x_k) \\ + g_{\{i,j\}}(x_i, x_j) + g_{\{i,k\}}(x_i, x_k) + g_{\{i,j\}}(x_j, x_k)\end{aligned}$$

- **Example**:

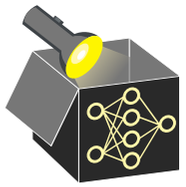
$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 - \sin(x_2 x_3) + 1$$

- **Note:** Again using centered PD-functions $\hat{f}^c_{S,PD}$ instead of components g_S
  $\rightsquigarrow$ things get complicated, e.g. for 3 features, definition becomes:

$$\begin{aligned}\hat{f}^c_{\{ijk\},PD}(x_i, x_j, x_k) =& \hat{f}^c_{\{ij\},PD}(x_i, x_j) + \hat{f}^c_{\{ik\},PD}(x_i, x_k) + \hat{f}^c_{\{jk\},PD}(x_j, x_k) \\ & - \hat{f}^c_{i,PD}(x_i) - \hat{f}^c_{j,PD}(x_j) - \hat{f}^c_{k,PD}(x_k)\end{aligned}$$

IDEA

POPESCU_2008



3-way interaction:

- **Definition:** $\hat{f}$ contains no 3-way interactions between features i, j, k, if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

$$\begin{aligned}\hat{f}_{\{ijk\},PD}(x_i, x_j, x_k) = g_\emptyset + g_i(x_i) + g_j(x_j) + g_k(x_k) \\ + g_{\{i,j\}}(x_i, x_j) + g_{\{i,k\}}(x_i, x_k) + g_{\{i,j\}}(x_j, x_k)\end{aligned}$$

- **Example**:

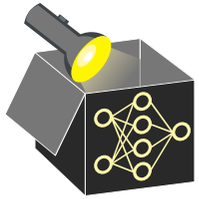
$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 - \sin(x_2 x_3) + 1$$

- **Note:** Again using centered PD-functions $\hat{f}^c_{S,PD}$ instead of components g_S
  $\rightsquigarrow$ things get complicated, e.g. for 3 features, definition becomes:

$$\begin{aligned}\hat{f}^c_{\{ijk\},PD}(x_i, x_j, x_k) =& \hat{f}^c_{\{ij\},PD}(x_i, x_j) + \hat{f}^c_{\{ik\},PD}(x_i, x_k) + \hat{f}^c_{\{jk\},PD}(x_j, x_k) \\ & - \hat{f}^c_{i,PD}(x_i) - \hat{f}^c_{j,PD}(x_j) - \hat{f}^c_{k,PD}(x_k)\end{aligned}$$

IDEA

Friedman and Popescu (2008)



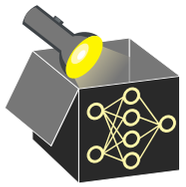
k-**way interaction:**

- **Analogous** for general *k*-way interactions between features $S = \{i_1, i_2, \ldots, i_k\}$: No *k*-way interaction, if

$$\hat{f}_{S,PD}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \sum_{\substack{V \subseteq S \\ |V|<k}} g_V(\mathbf{x}_V)$$

IDEA

POPESCU_2008



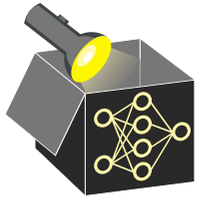
k-**way interaction:**

- **Analogous** for *k*-way interactions between feat $S = \{i_1, i_2, \ldots, i_k\}$: No *k*-way interaction, if

$$\hat{f}_{S,PD}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \sum_{\substack{V \subseteq S \\ |V|<k}} g_V(\mathbf{x}_V)$$

IDEA

Friedman and Popescu (2008)



k-**way interaction:**

- **Analogous** for general k-way interactions between features $S = \{i_1, i_2, \ldots, i_k\}$: No k-way interaction, if

$$\hat{f}_{S,PD}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \sum_{\substack{V \subseteq S \\ |V|<k}} g_V(\mathbf{x}_V)$$

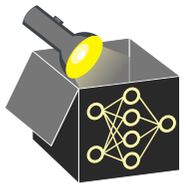
Overall interaction:

- Question: Does feature j interact with any other feature at all?

$\Rightarrow$ H-statistic analogous to 2-way interactions, but for feature sets $S = \{j\}$ and $-S = \{1, \ldots, p\} \setminus \{j\}$ instead of two single features:

IDEA

POPESCU_2008



k-**way interaction:**

- **Analogous** for k-way interactions between feat $S = \{i_1, i_2, \ldots, i_k\}$: No k-way interaction, if

$$\hat{f}_{S,PD}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \sum_{\substack{V \subseteq S \\ |V|<k}} g_V(\mathbf{x}_V)$$

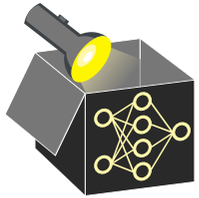
Overall interaction:

- Question: Does feature j interact with any other feature at all?

$\Rightarrow$ H-statistic analogous to 2-way interactions, but for feature sets $S = \{j\}$ and $-S = \{1, \ldots, p\} \setminus \{j\}$ instead of two single features:

IDEA

Friedman and Popescu (2008)



k-**way interaction:**

- **Analogous** for general *k*-way interactions between features $S = \{i_1, i_2, \ldots, i_k\}$: No *k*-way interaction, if

$$\hat{f}_{S,PD}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \sum_{\substack{V \subseteq S \\ |V|<k}} g_V(\mathbf{x}_V)$$

Overall interaction:

- Question: Does feature j interact with any other feature at all?

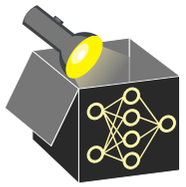
$\Rightarrow$ H-statistic analogous to 2-way interactions, but for feature sets $S = \{j\}$ and $-S = \{1, \ldots, p\} \setminus \{j\}$ instead of two single features:

$$\hat{f}(\mathbf{x}) - g_\emptyset = \hat{f}^c_{\{1,\ldots,p\},PD}(\mathbf{x}) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{-j,PD}(\mathbf{x}_{-j}) = \sum_{\substack{S: j \in S \\ |S| \geq 2}} g_S(\mathbf{x}_S)$$

- $-j$ denotes $-S = \{1, \ldots, p\} \setminus \{j\}$, i.e. all other features
- $\hat{f}_{-j,PD}(\mathbf{x}_{-j})$: $(p-1)$-dim PD function of all p features except feature j

IDEA

POPESCU_2008



k-**way interaction:**

- **Analogous** for *k*-way interactions between feat $S = \{i_1, i_2, \ldots, i_k\}$: No *k*-way interaction, if

$$\hat{f}_{S,PD}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \sum_{\substack{V \subseteq S \\ |V|<k}} g_V(\mathbf{x}_V)$$

Overall interaction:

- Question: Does feature j interact with any other feature at all?

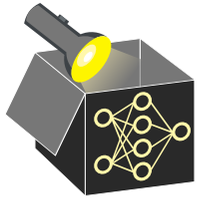
$\Rightarrow$ H-statistic analogous to 2-way interactions, but for feature sets $S = \{j\}$ and $-S = \{1, \ldots, p\} \setminus \{j\}$ instead of two single features:

$$\hat{f}(\mathbf{x}) - g_\emptyset = \hat{f}^c_{\{1,\ldots,p\},PD}(\mathbf{x}) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{-j,PD}(\mathbf{x}_{-j}) = \sum_{\substack{S: j \in S \\ |S| \geq 2}} g_S(\mathbf{x}_S)$$

- $-j$ denotes $-S = \{1, \ldots, p\} \setminus \{j\}$, i.e. all other features
- $\hat{f}_{-j,PD}(\mathbf{x}_{-j})$: $(p-1)$-dim PD function of all p features except feature j

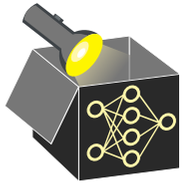
2-WAY INTERACTION STRENGTH

- **Question:** How to measure interaction strength without computing functional decomposition components g_S?

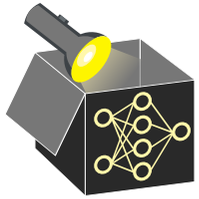


2-WAY INTERACTION STRENGTH

- **Question:** How to measure interaction strength without computing functional decomposition components g_S?



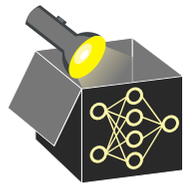
2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

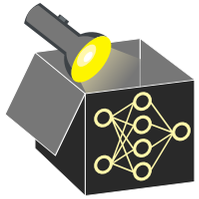
2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

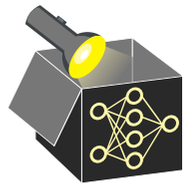
$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

- **H-statistic** for 2-way interaction between feature j and k:

$$H^2_{jk} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{k,PD}(x_k^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)})\right)^2}$$

2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

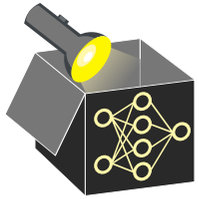
$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

- **H-statistic** for 2-way interaction between feature j and k:

$$H^2_{jk} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{k,PD}(x_k^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)})\right)^2}$$

2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

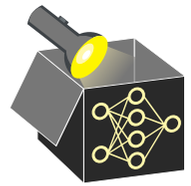
- **H-statistic** for 2-way interaction between feature j and k:

$$H^2_{jk} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{k,PD}(x_k^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)})\right)^2}$$

$\Rightarrow$ H^2_{jk} measures strength of this interaction quantitatively
H^2_{jk} small (close to 0) for weak interaction, close to 1 for strong interaction

2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

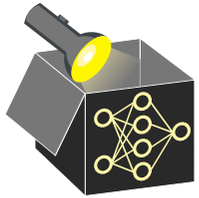
- **H-statistic** for 2-way interaction between feature j and k:

$$H^2_{jk} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{k,PD}(x_k^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)})\right)^2}$$

$\Rightarrow$ H^2_{jk} measures strength of this interaction quantitatively
H^2_{jk} small (close to 0) for weak interaction, close to 1 for strong interaction

2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

- **H-statistic** for 2-way interaction between feature j and k:

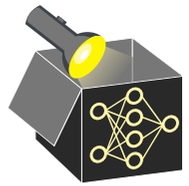
$$H^2_{jk} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{k,PD}(x_k^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)})\right)^2}$$

$\Rightarrow$ H^2_{jk} measures strength of this interaction quantitatively
H^2_{jk} small (close to 0) for weak interaction, close to 1 for strong interaction

- **Note:** Again, definition also usable without any probabilities or data distribution

2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \; ?$$

- **H-statistic** for 2-way interaction between feature j and k:

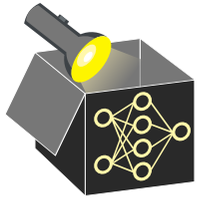
$$H^2_{jk} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{k,PD}(x_k^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)})\right)^2}$$

$\Rightarrow$ H^2_{jk} measures strength of this interaction quantitatively
H^2_{jk} small (close to 0) for weak interaction, close to 1 for strong interaction

- **Note:** Again, definition also usable without probabilities or data distrib.

H-STATISTIC: EXAMPLES



Note: Again, definition also usable without any probability or data distribution

Example

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \quad (x_1, x_2) \in [0, 1]^2$$
$$\hat{f}^c_{1,PD}(x_1) = -2x_1 + 0.5|x_1| + 0.75$$
$$\hat{f}^c_{2,PD}(x_2) = 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05$$
$$\hat{f}^c_{1,2;PD}(x_1, x_2) = 1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e$$

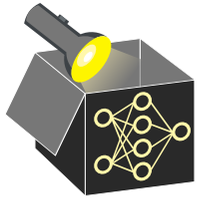
H-STATISTIC: EXAMPLES

Note: Again, definition also usable without any probability or data distribution

Example

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \quad (x_1, x_2) \in [0, 1]^2$$
$$\hat{f}^c_{1,PD}(x_1) = -2x_1 + 0.5|x_1| + 0.75$$
$$\hat{f}^c_{2,PD}(x_2) = 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05$$
$$\hat{f}^c_{1,2;PD}(x_1, x_2) = 1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e$$

H-STATISTIC: EXAMPLES

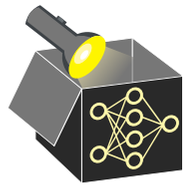


Note: Again, definition also usable without any probability or data distribution

Example

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \quad (x_1, x_2) \in [0,1]^2$$
$$\hat{f}^c_{1,PD}(x_1) = -2x_1 + 0.5|x_1| + 0.75$$
$$\hat{f}^c_{2,PD}(x_2) = 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05$$
$$\hat{f}^c_{1,2;PD}(x_1, x_2) = 1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e$$
$$\implies H^2_{12} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$
$$= \frac{\mathbb{E}\left[\left(|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25\right)^2\right]}{\mathbb{E}\left[\left(1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e\right)^2\right]} > 0$$

H-STATISTIC: EXAMPLES

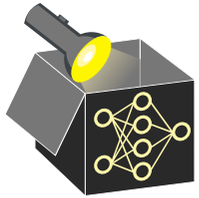


Note: Again, definition also usable without any probability or data distribution

Example

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \quad (x_1, x_2) \in [0,1]^2$$
$$\hat{f}^c_{1,PD}(x_1) = -2x_1 + 0.5|x_1| + 0.75$$
$$\hat{f}^c_{2,PD}(x_2) = 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05$$
$$\hat{f}^c_{1,2;PD}(x_1, x_2) = 1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e$$
$$\implies H^2_{12} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$
$$= \frac{\mathbb{E}\left[\left(|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25\right)^2\right]}{\mathbb{E}\left[\left(1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e\right)^2\right]} > 0$$

3-WAY INTERACTION STRENGTH



- Same idea as for 2-way, but different formula (see before):

$$\hat{f}^c_{\{ijk\},PD}(x_i, x_j, x_k) = \hat{f}^c_{\{ij\},PD}(x_i, x_j) + \hat{f}^c_{\{ik\},PD}(x_i, x_k) + \hat{f}^c_{\{jk\},PD}(x_j, x_k) - \hat{f}^c_{i,PD}(x_i) - \hat{f}^c_{j,PD}(x_j) - \hat{f}^c_{k,PD}(x_k)$$

$\Rightarrow$ H-statistic for a 3-way interaction between features i, j and k:

$$H^2_{ijk} = \frac{\text{Var}\left[\hat{f}^c_{ijk,PD}(X_i, X_j, X_k) - \hat{f}^c_{ij,PD}(X_i, X_j) - \hat{f}^c_{ik,PD}(X_i, X_k) - \hat{f}^c_{jk,PD}(X_j, X_k) + \hat{f}^c_{i,PD}(X_i) + \hat{f}^c_{j,PD}(X_j) + \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{ijk,PD}(X_i, X_j, X_k)\right]}$$

- Analogous for higher order interactions, but more complicated

3-WAY INTERACTION STRENGTH

- Same idea as for 2-way, but different formula (see before):

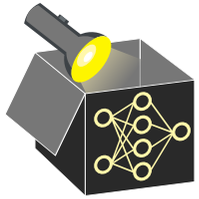
$$\hat{f}^c_{\{ijk\},PD}(x_i, x_j, x_k) = \hat{f}^c_{\{ij\},PD}(x_i, x_j) + \hat{f}^c_{\{ik\},PD}(x_i, x_k) + \hat{f}^c_{\{jk\},PD}(x_j, x_k) - \hat{f}^c_{i,PD}(x_i) - \hat{f}^c_{j,PD}(x_j) - \hat{f}^c_{k,PD}(x_k)$$

$\Rightarrow$ H-statistic for a 3-way interaction between features i, j and k:

$$H^2_{ijk} = \frac{\text{Var}\left[\hat{f}^c_{ijk,PD}(X_i, X_j, X_k) - \hat{f}^c_{ij,PD}(X_i, X_j) - \hat{f}^c_{ik,PD}(X_i, X_k) - \hat{f}^c_{jk,PD}(X_j, X_k) + \hat{f}^c_{i,PD}(X_i) + \hat{f}^c_{j,PD}(X_j) + \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{ijk,PD}(X_i, X_j, X_k)\right]}$$

- Analogous for higher order interactions, but more complicated

OVERALL INTERACTION STRENGTH



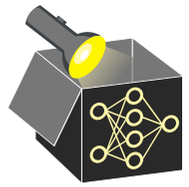
- Measure overall strength of interactions between feature j and all other features

$\Rightarrow$ **H-statistic** analogous to 2-way interaction:

$$H_j^2 = \frac{\text{Var}\left[\hat{f}^c(\mathbf{X}) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{-j,PD}(\mathbf{X}_{-j})\right]}{\text{Var}\left[\hat{f}^c(\mathbf{X})\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{-j,PD}(\mathbf{x}^{(i)}_{-j})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)})\right)^2}$$

OVERALL INTERACTION STRENGTH



- Measure overall strength of interactions between feat j and all other feats

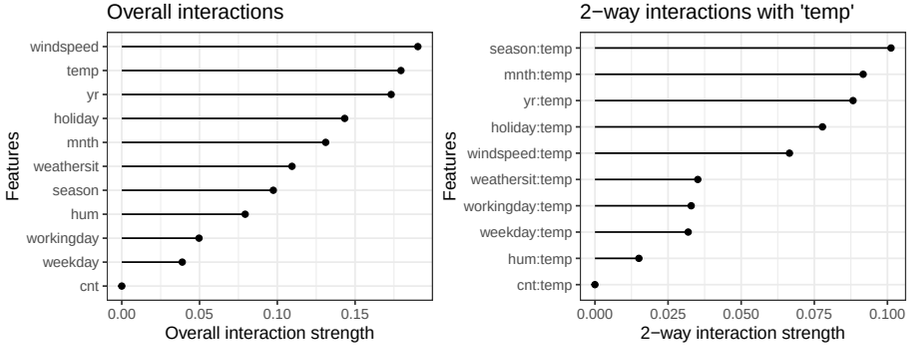
$\Rightarrow$ **H-statistic** analogous to 2-way interaction:

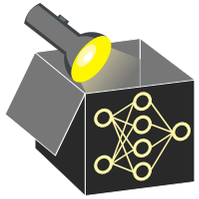
$$H_j^2 = \frac{\text{Var}\left[\hat{f}^c() - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{-j,PD}(_{-j})\right]}{\text{Var}\left[\hat{f}^c()\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{-j,PD}(\mathbf{x}^{(i)}_{-j})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)})\right)^2}$$

H-STATISTIC: EXAMPLE

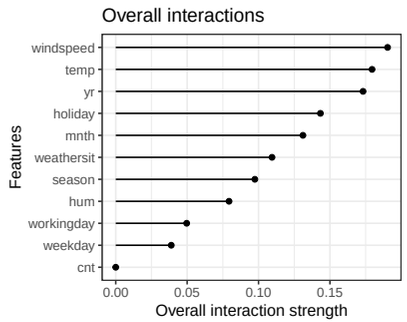
Measure interactions of a random forest for the bike data set

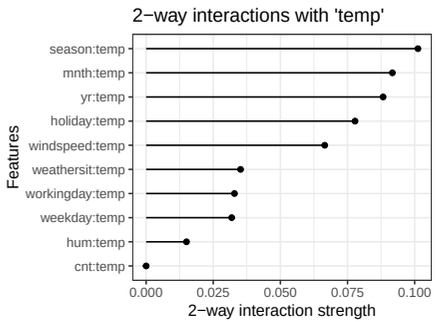


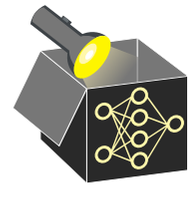


H-STATISTIC: EXAMPLE

Measure interactions of a random forest for the bike data set

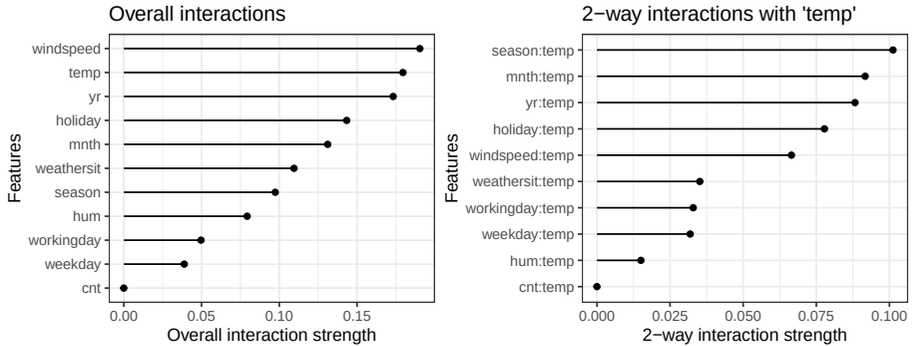


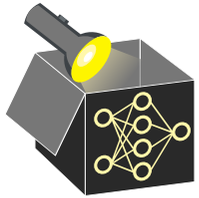




H-STATISTIC: EXAMPLE

Measure interactions of a random forest for the bike data set



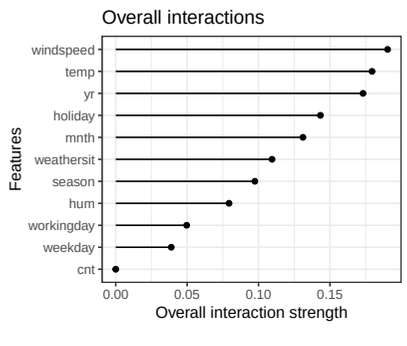


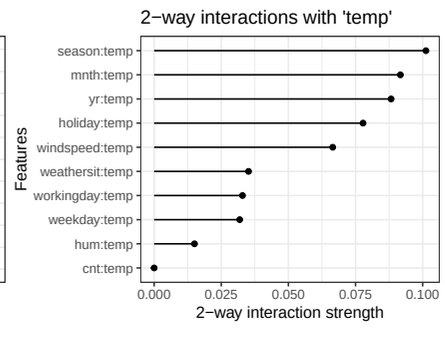
Remarks and Conclusion:

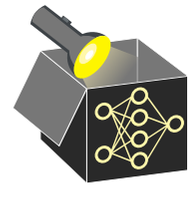
- H-statistic provides **general definition of interactions** + an **algorithm for computation**
  Also adjustable to categorical / discrete features and / or function values
- For interaction order k still needs ¸ $\approx 2^k$ PD-functions
- Statistical test for whether interactions are present using this statistic

H-STATISTIC: EXAMPLE

Measure interactions of a random forest for the bike data set





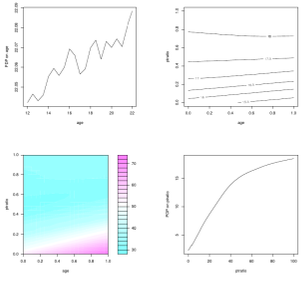


Remarks and Conclusion:

- H-statistic provides **general definition of interactions** + an **algorithm for computation**
  Also adjustable to categorical / discrete features and / or function values
- For interaction order k still needs ¸ $\approx 2^k$ PD-functions
- Statistical test for whether interactions are present using this statistic

Interpretable Machine Learning

Theory of Standard fANOVA

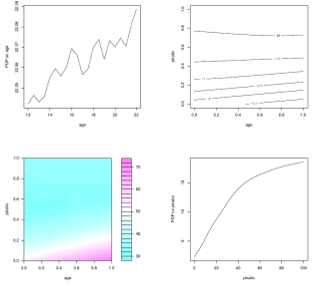


Learning goals

- Properties of classical fANOVA, reason for its popularity
- Equivalent definition of classical fANOVA
- Understand the role constraints play for any functional decomposition

Interpretable Machine Learning

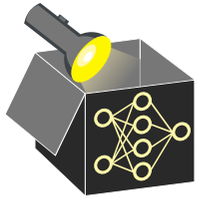
Functional Decompositions
Theory of Standard fANOVA



Learning goals

- Properties of classical fANOVA, reason for its popularity
- Equivalent definition of classical fANOVA
- Understand the role constraints play for any functional decomposition

EXAMPLE: FANOVA ALGORITHM



- Remember: Functional decomposition in general not unique
- **Standard fANOVA** only one possible approach
- Example:

$$\begin{aligned}\hat{f}(x_1, x_2) &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \\ &= \underbrace{2.95 + 0.3e.}_{g_\emptyset} + \underbrace{-2x_1 + 0.5|x_1| + 0.75}_{g_1(x_1)} \\ &\quad + \underbrace{0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05}_{g_2(x_2)} + \underbrace{|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25}_{g_{1,2}(x_1,x_2)}\end{aligned}$$

$\rightsquigarrow$ seems arbitrarily chosen?
$\longleftrightarrow$ Show: Standard fANOVA fulfills specific desirable properties or **constraints**

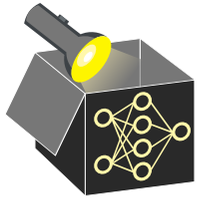
EXAMPLE: FANOVA ALGORITHM

- Remember: Functional decomposition in general not unique
- **Standard fANOVA** only one possible approach
- Example:

$$\begin{aligned}\hat{f}(x_1, x_2) &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \\ &= \underbrace{2.95 + 0.3e.}_{g_\emptyset} + \underbrace{-2x_1 + 0.5|x_1| + 0.75}_{g_1(x_1)} \\ &\quad + \underbrace{0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05}_{g_2(x_2)} + \underbrace{|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25}_{g_{1,2}(x_1,x_2)}\end{aligned}$$

$\rightsquigarrow$ seems arbitrarily chosen?
$\longleftrightarrow$ Show: Standard fANOVA fulfills specific desirable properties or **constraints**

CONSTRAINTS FOR STANDARD FANOVA ALGORITHM

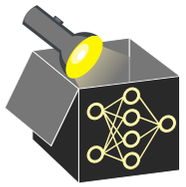


Theorem

Features independent $\implies$ *The components defined by standard fANOVA fulfill the so-called vanishing conditions:*

$$\mathbb{E}_{X_j}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0 \quad \textit{for any } j \in S \textit{ and } S \subseteq \{1, \ldots, p\}$$

CONSTRAINTS FOR STANDARD FANOVA ALGO.

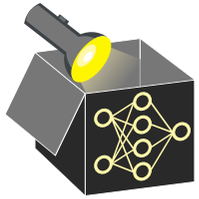


Theorem

Features independent $\implies$ *The components defined by standard fANOVA fulfill the so-called vanishing conditions:*

$$\mathbb{E}_{X_j}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0 \quad \textit{for any } j \in S \textit{ and } S \subseteq \{1, \ldots, p\}$$

CONSTRAINTS FOR STANDARD FANOVA ALGORITHM



Theorem

Features independent $\implies$ *The components defined by standard fANOVA fulfill the so-called vanishing conditions:*

$$\mathbb{E}_{X_j}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0 \quad \textit{for any } j \in S \textit{ and } S \subseteq \{1, \ldots, p\}$$

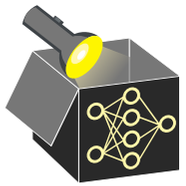
Implications:

- For any component g_S, all its PD-functions are 0:

$$\mathbb{E}_{X_V}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(\mathbf{x}_V) = 0 \quad \text{for any } V \subsetneq S \text{ and } S \subseteq \{1, \ldots, p\}$$

  $\rightsquigarrow$ g_S contains no lower-order effects, but only pure interaction term (compare H-statistic)

CONSTRAINTS FOR STANDARD FANOVA ALGO.



Theorem

Features independent $\implies$ *The components defined by standard fANOVA fulfill the so-called vanishing conditions:*

$$\mathbb{E}_{X_j}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0 \quad \textit{for any } j \in S \textit{ and } S \subseteq \{1, \ldots, p\}$$

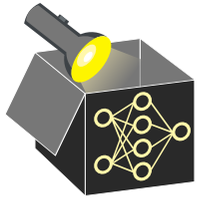
Implications:

- For any component g_S, all its PD-functions are 0:

$$\mathbb{E}_{X_V}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(\mathbf{x}_V) = 0 \quad \text{for any } V \subsetneq S \text{ and } S \subseteq \{1, \ldots, p\}$$

  $\rightsquigarrow$ g_S contains no lower-order effects, but only pure interaction term (compare H-statistic)

CONSTRAINTS FOR STANDARD FANOVA ALGORITHM



Theorem

Features independent $\implies$ *The components defined by standard fANOVA fulfill the so-called vanishing conditions:*

$$\mathbb{E}_{X_j}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0 \quad \textit{for any } j \in S \textit{ and } S \subseteq \{1, \ldots, p\}$$

Implications:

- For any component g_S, all its PD-functions are 0:

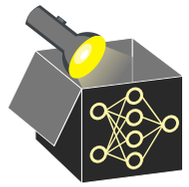
  $$\mathbb{E}_{X_V}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(\mathbf{x}_V) = 0 \quad \text{for any } V \subsetneq S \text{ and } S \subseteq \{1, \ldots, p\}$$

  $\rightsquigarrow$ g_S contains no lower-order effects, but only pure interaction term (compare H-statistic)
- All components are orthogonal, i.e., mutually independent and uncorrelated:

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

- This implies variance decomposition used to define Sobol indices:
  $\text{Var}[\hat{f}(\mathbf{x})] = \sum_{S \subseteq \{1,\ldots,p\}} \text{Var}\left[g_S(\mathbf{x}_S)\right]$

CONSTRAINTS FOR STANDARD FANOVA ALGO.



Theorem

Features independent $\implies$ *The components defined by standard fANOVA fulfill the so-called vanishing conditions:*

$$\mathbb{E}_{X_j}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0 \quad \textit{for any } j \in S \textit{ and } S \subseteq \{1, \ldots, p\}$$

Implications:

- For any component g_S, all its PD-functions are 0:

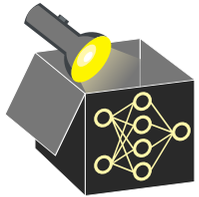
  $$\mathbb{E}_{X_V}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(\mathbf{x}_V) = 0 \quad \text{for any } V \subsetneq S \text{ and } S \subseteq \{1, \ldots, p\}$$

  $\rightsquigarrow$ g_S contains no lower-order effects, but only pure interaction term (compare H-statistic)
- All components are orthogonal, i.e., mutually indep. and uncorrelated:

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

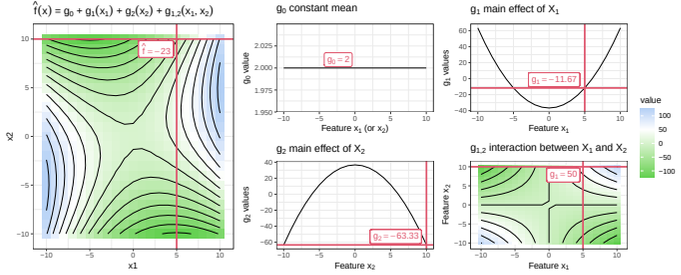
- This implies variance decomposition used to define Sobol indices:
  $\text{Var}[\hat{f}(\mathbf{x})] = \sum_{S \subseteq \{1,\ldots,p\}} \text{Var}\left[g_S(\mathbf{x}_S)\right]$

EXAMPLES REVISITED



Example: $\hat{f}(\mathbf{x}) = 2 + x_1^2 - x_2^2 + x_1 \cdot x_2$ (e.g., for $x_1 = 5$ and $x_2 = 10$ we have $\hat{f}(\mathbf{x}) = -23$)

- Computation of components using feature values $x_1 = x_2 = (-10, -9, \ldots, 10)^\top$ gives:



For $x_1 = 5$ and $x_2 = 10$:

- $g_\emptyset = 2$
- $g_1(x_1) = -9.67$
- $g_2(x_2) = -65.33$
- $g_{1,2}(x_1, x_2) = 50$

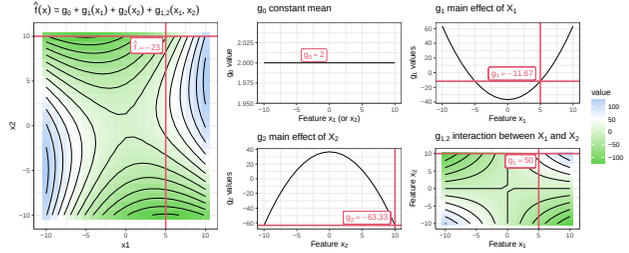
$\Rightarrow \hat{f}(\mathbf{x}) = -23$

- Vanishing condition means:
  - g_1 and g_2 are mean-centered w.r.t. marginal distribution of x_1 and x_2
  - Integral of $g_{1,2}$ over marginal distribution x_1 (or x_2) is always 0.

EXAMPLES REVISITED

Example: $\hat{f}(\mathbf{x}) = 2 + x_1^2 - x_2^2 + x_1 \cdot x_2$ (e.g., for $x_1 = 5$ and $x_2 = 10$ we have $\hat{f}(\mathbf{x}) = -23$)

- Computation of components using feature values $x_1 = x_2 = (-10, -9, \ldots, 10)^\top$ gives:



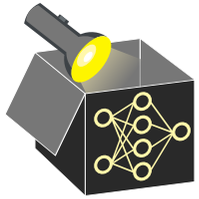
For $x_1 = 5$ and $x_2 = 10$:

- $g_\emptyset = 2$
- $g_1(x_1) = -9.67$
- $g_2(x_2) = -65.33$
- $g_{1,2}(x_1, x_2) = 50$

$\Rightarrow \hat{f}(\mathbf{x}) = -23$

- Vanishing condition means:
  - g_1 and g_2 are mean-centered w.r.t. marginal distribution of x_1 and x_2
  - Integral of $g_{1,2}$ over marginal distribution x_1 (or x_2) is always 0.

EXAMPLES REVISITED



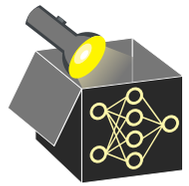
Example

$$\begin{aligned}\hat{f}(x_1, x_2) &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \\ &= \underbrace{2.95 + 0.3e.}_{g_\emptyset} + \underbrace{-2x_1 + 0.5|x_1| + 0.75}_{g_1(x_1)} \\ &\quad + \underbrace{0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05}_{g_2(x_2)} + \underbrace{|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25}_{g_{1,2}(x_1,x_2)}\end{aligned}$$

$\Longrightarrow$ Main effect terms inside $g_{1,2}$ are chosen exactly such that the one-dimensional PDPs of $g_{1,2}$ vanish

$\Longrightarrow$ Same for constant terms inside g_1 and g_2: Ensure centering

EXAMPLES REVISITED



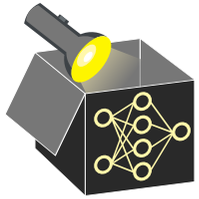
Example

$$\begin{aligned}\hat{f}(x_1, x_2) &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \\ &= \underbrace{2.95 + 0.3e.}_{g_\emptyset} + \underbrace{-2x_1 + 0.5|x_1| + 0.75}_{g_1(x_1)} \\ &\quad + \underbrace{0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05}_{g_2(x_2)} + \underbrace{|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25}_{g_{1,2}(x_1,x_2)}\end{aligned}$$

$\Longrightarrow$ Main effect terms inside $g_{1,2}$ are chosen exactly such that the one-dimensional PDPs of $g_{1,2}$ vanish

$\Longrightarrow$ Same for constant terms inside g_1 and g_2: Ensure centering

EXAMPLES REVISITED



Example

$$\begin{aligned}\hat{f}(x_1, x_2) &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \\ &= \underbrace{2.95 + 0.3e.}_{g_\emptyset} + \underbrace{-2x_1 + 0.5|x_1| + 0.75}_{g_1(x_1)} \\ &\quad + \underbrace{0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05}_{g_2(x_2)} + \underbrace{|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25}_{g_{1,2}(x_1,x_2)}\end{aligned}$$

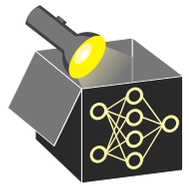
$\Longrightarrow$ Main effect terms inside $g_{1,2}$ are chosen exactly such that the one-dimensional PDPs of $g_{1,2}$ vanish

$\Longrightarrow$ Same for constant terms inside g_1 and g_2: Ensure centering

Example

From in-class exercise: $g(x_1, x_2) = \beta_{12}\,(x_1 - \mu_1)(x_2 - \mu_2)$

EXAMPLES REVISITED



Example

$$\begin{aligned}\hat{f}(x_1, x_2) &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \\ &= \underbrace{2.95 + 0.3e.}_{g_\emptyset} + \underbrace{-2x_1 + 0.5|x_1| + 0.75}_{g_1(x_1)} \\ &\quad + \underbrace{0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05}_{g_2(x_2)} + \underbrace{|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25}_{g_{1,2}(x_1,x_2)}\end{aligned}$$

$\Longrightarrow$ Main effect terms inside $g_{1,2}$ are chosen exactly such that the one-dimensional PDPs of $g_{1,2}$ vanish

$\Longrightarrow$ Same for constant terms inside g_1 and g_2: Ensure centering

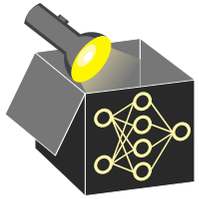
Example

From in-class exercise: $g(x_1, x_2) = \beta_{12}\,(x_1 - \mu_1)(x_2 - \mu_2)$

Interpretable Machine Learning – 4 / 5

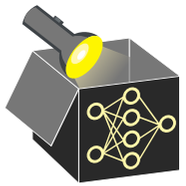
CONSTRAINTS: EQUIVALENT CHARACTERIZATION

- So far: Definition of standard fANOVA implies vanishing conditions



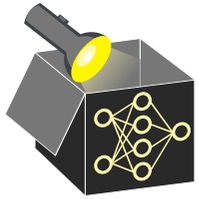
CONSTRAINTS: EQUIVALENT CHARACTERIZATION

- So far: Definition of standard fANOVA implies vanishing conditions



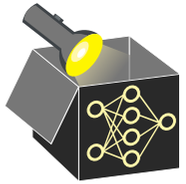
CONSTRAINTS: EQUIVALENT CHARACTERIZATION

- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.

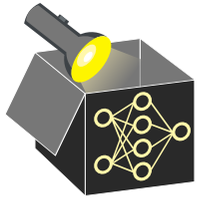


CONSTRAINTS: EQUIVALENT CHARACTERIZATION

- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.

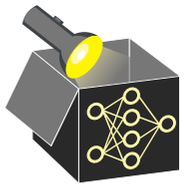


CONSTRAINTS: EQUIVALENT CHARACTERIZATION



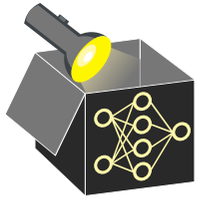
- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.
- In other words: Vanishing conditions are equivalent characterization

CONSTRAINTS: EQUIVALENT CHARACTERIZATION



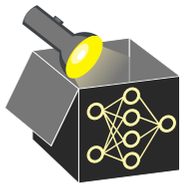
- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.
- In other words: Vanishing conditions are equivalent characterization

CONSTRAINTS: EQUIVALENT CHARACTERIZATION



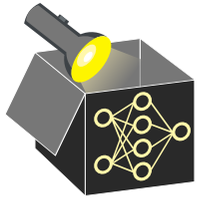
- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.
- In other words: Vanishing conditions are equivalent characterization
- In general: Functional decompositions can be defined by sets of constraints

CONSTRAINTS: EQUIVALENT CHARACTERIZATION



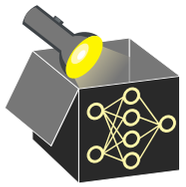
- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.
- In other words: Vanishing conditions are equivalent characterization
- In general: Functional decomp. can be defined by sets of constraints

CONSTRAINTS: EQUIVALENT CHARACTERIZATION



- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.
- In other words: Vanishing conditions are equivalent characterization
- In general: Functional decompositions can be defined by sets of constraints
- Many other methods to compute decompositions exist, each with their set of constraints

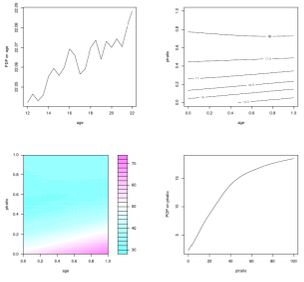
CONSTRAINTS: EQUIVALENT CHARACTERIZATION



- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\implies$ Any functional decomposition fulfilling the vanishing conditions must be the standard fANOVA decomposition.
- In other words: Vanishing conditions are equivalent characterization
- In general: Functional decomp. can be defined by sets of constraints
- Many other methods to compute decompositions exist, each with their set of constraints

Interpretable Machine Learning

Functional Decompositions: Further Methods

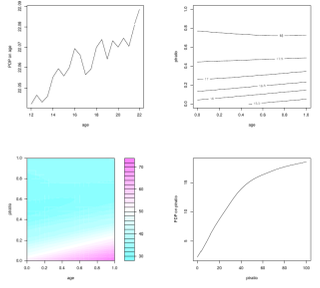


Learning goals

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

Interpretable Machine Learning

Functional Decompositions
Further Methods

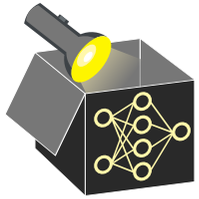


Learning goals

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

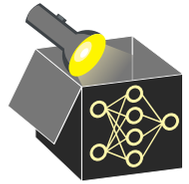
LIMITATIONS OF CLASSICAL FANOVA

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

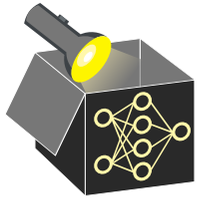


LIMITATIONS OF CLASSICAL FANOVA

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

Example

Assume dependency $2x_1^2 = x_2$ and

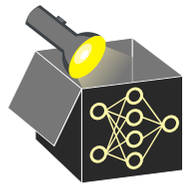
$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_2x_3 + 1.$$

$\rightsquigarrow$ 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 $\implies$ Standard fANOVA does NOT fulfill vanishing conditions

Example

Assume dependency $2x_1^2 = x_2$ and

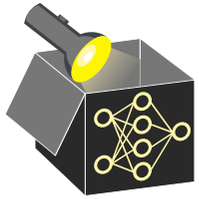
$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_2x_3 + 1.$$

$\rightsquigarrow$ 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 $\implies$ Standard fANOVA does NOT fulfill vanishing conditions

Example

Assume dependency $2x_1^2 = x_2$ and

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

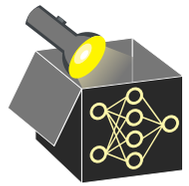
$\rightsquigarrow$ Following two decompositions would both "make sense":

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

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

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

LIMITATIONS OF CLASSICAL FANOVA



- Standard fANOVA builds on PD-functions
- *Remember:* Problems of PDPs for **correlated / dependent features**
- Here: Dependent features $\implies$ 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":

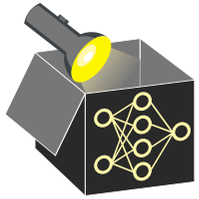
$$\hat{f}(x_1, x_2, x_3) = \underbrace{1}_{g_\emptyset} + \underbrace{(-2x_1)}_{g_1(x_1)} + \underbrace{(-2\sin(x_3))}_{g_3(x_3)} + \underbrace{|x_1|x_2}_{g_{1,2}(x_1,x_2)} + \underbrace{0.5x_2x_3}_{g_{2,3}(x_2,x_3)}$$

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

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

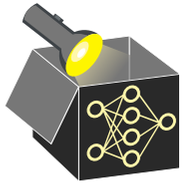
ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA

- Algorithm proposed by ▸ Hooker (2007)
- Generalizes standard fANOVA to situations with dependent features

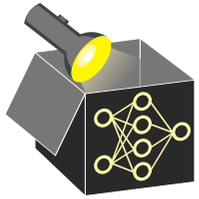


ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA

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



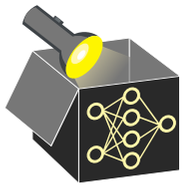
ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA



- 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”:

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

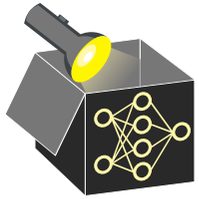
ALTERNATIVE: GENERALIZED FUNCTIONAL ANOVA



- 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”:

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

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":

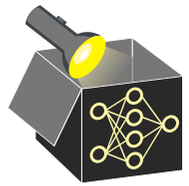
$$\mathbb{E}_{\mathbf{x}}\left[g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)\right] = 0 \quad \forall V \subsetneq S$$

  $\rightsquigarrow$ 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 decomposition 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":

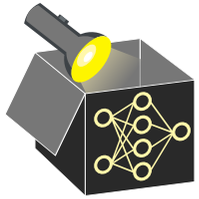
$$\mathbb{E}\left[g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)\right] = 0 \quad \forall V \subsetneq S$$

  $\rightsquigarrow$ 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”:

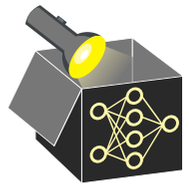
$$\mathbb{E}_{\mathbf{x}}\left[g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)\right] = 0 \quad \forall V \subsetneqq S$$

  $\rightsquigarrow$ 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 decomposition for arbitrary settings

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

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”:

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

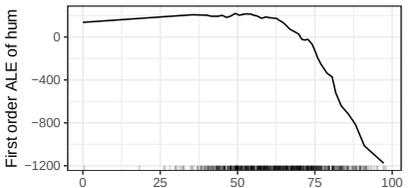
  $\rightsquigarrow$ Only components are orthogonal where $g_V(\mathbf{x}_V)$ is “lower in hierarchy” than $g_S(\mathbf{x}_S)$

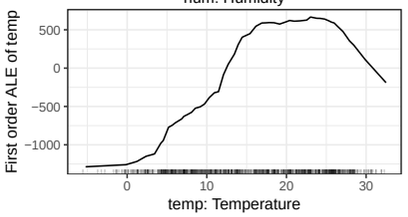
$\Longrightarrow$ Generalized fANOVA provides functional decomp. for arbitrary settings

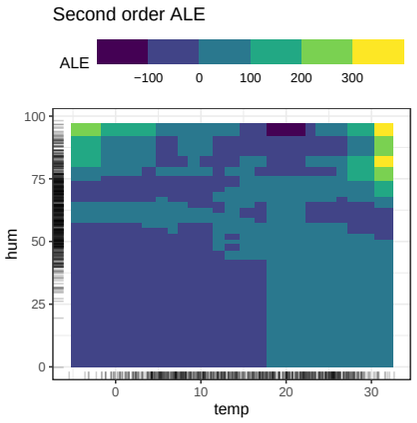
- **Advantage:** Also provides a variance decomposition
- **Problems:**
- Difficult to estimate, involves manual choice of a “weight function”
- Computationally very costly

REVISITING ALE PLOTS

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

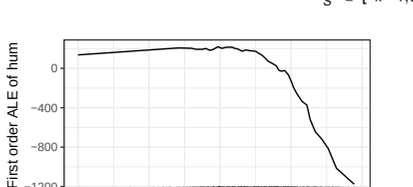


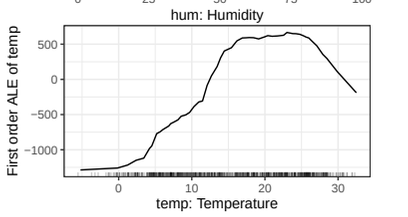


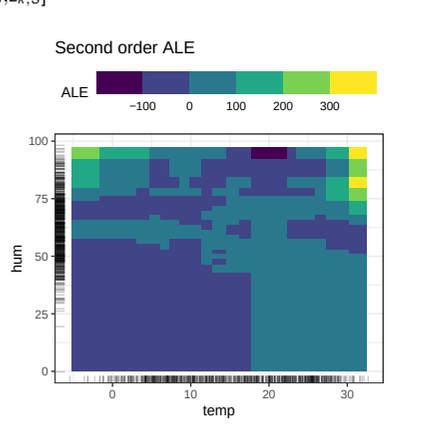


REVISITING ALE PLOTS

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



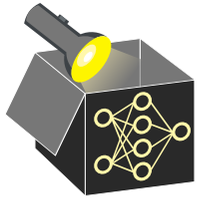




ALE DECOMPOSITION

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

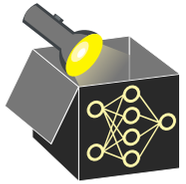
$\rightarrow$ Gives full functional decomposition of ALE plots



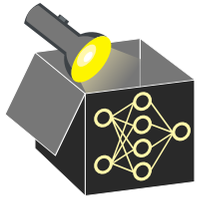
ALE DECOMPOSITION

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

$\rightarrow$ Gives full functional decomposition of ALE plots



ALE DECOMPOSITION



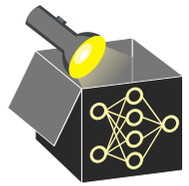
- One can define ALE plots for arbitrary many variables (similar to PDPs vs. PD-functions)

$\rightarrow$ Gives full functional decomposition of ALE plots

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

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

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

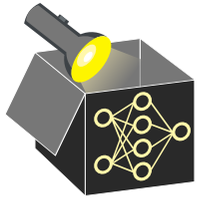
- **Advantages:** Handle dependencies well + computationally fast
- Constraints / orthogonality properties more complicated

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

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?

- If computed, offer a lot of insight into a model or function, i.p. high-dimensional

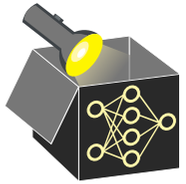
$\rightarrow$ Complete analysis of all interactions



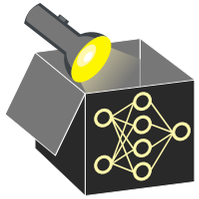
CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?

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

$\rightarrow$ Complete analysis of all interactions



CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?

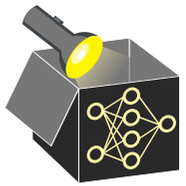


- If computed, offer a lot of insight into a model or function, i.p. high-dimensional

$\rightarrow$ Complete analysis of all interactions

- Very important theoretical concept:
  - Theoretical framework for general definition of interactions (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?

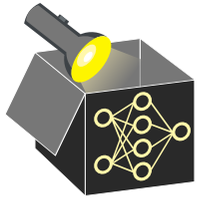


- If computed, offer many insights into a model / function, i.p. high-dim.

$\rightarrow$ 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)

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



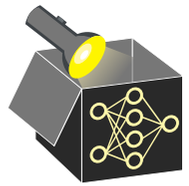
- If computed, offer a lot of insight into a model or function, i.p. high-dimensional

$\rightarrow$ Complete analysis of all interactions

- Very important theoretical concept:
  - Theoretical framework for general definition of interactions (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?



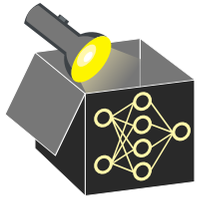
- If computed, offer many insights into a model / function, i.p. high-dim.

$\rightarrow$ 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)

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



- If computed, offer a lot of insight into a model or function, i.p. high-dimensional

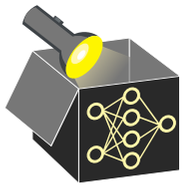
$\rightarrow$ Complete analysis of all interactions

- Very important theoretical concept:
  - Theoretical framework for general definition of interactions (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, evtl. infeasible
  - ALE: No variance decomposition

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



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

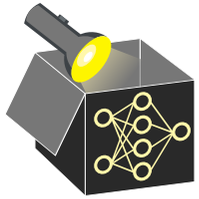
$\rightarrow$ Complete analysis of all interactions

- 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

CONCLUSION: HOW USEFUL ARE FUNCTIONAL DECOMPOSITIONS?



- If computed, offer a lot of insight into a model or function, i.p. high-dimensional

$\rightarrow$ Complete analysis of all interactions

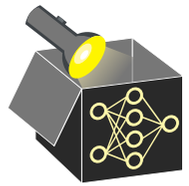
- Very important theoretical concept:
  - Theoretical framework for general definition of interactions (H-statistic)
  - Theoretical background for many IML methods: GAMs and EBMs, ICE, PDPs and PD-functions, ALE plots, Shapley values, Feature importance methods (see later)
- In practice often infeasible (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, evtl. infeasible
  - ALE: No variance decomposition

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

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