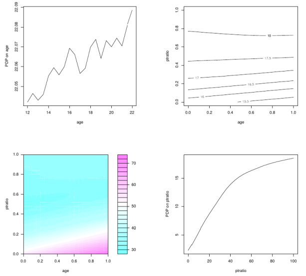
Interpretable Machine Learning

Theory of Standard fANOVA

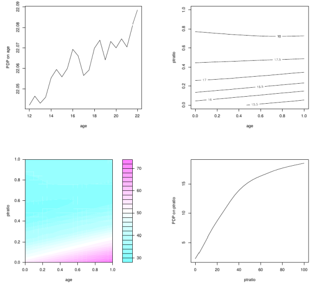


Learning goals

- 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

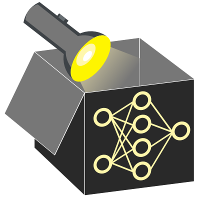
Functional Decompositions
Theory of Standard fANOVA



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

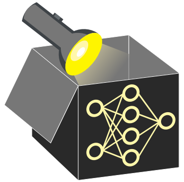


- Remember: Functional decomposition in general not unique
- **Standard fANOVA** only one possible approach
- Example:

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

$\leadsto$ 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:

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

$\leadsto$ 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:

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

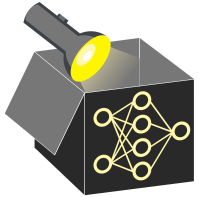
CONSTRAINTS FOR STANDARD FANOVA ALGO.

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:

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

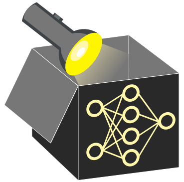
Implications:

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

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

  $\rightsquigarrow$ 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:

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

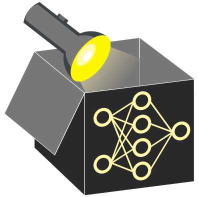
Implications:

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

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

  $\rightsquigarrow$ 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:

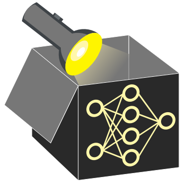
$$\mathbb{E}_{x_V}\left[g_S(\mathbf{x}_S)\right] = \int g_S(\mathbf{x}_S) d\mathbb{P}(\mathbf{x}_V) = 0 \quad \text{for any } V \subsetneq S \text{ and } S \subseteq \{1, \ldots, p\}$$

  $\rightsquigarrow$ 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:

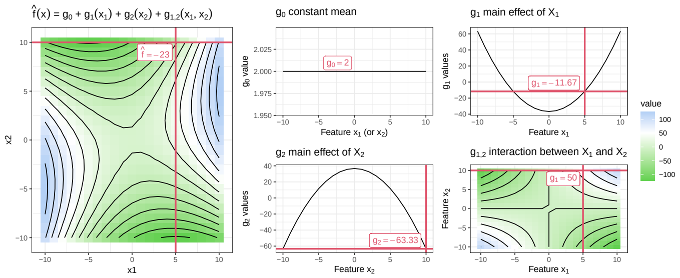
$$\forall V \neq S: \quad \mathbb{E}\left[g_V(\mathbf{x}_V) g_S(\mathbf{x}_S)\right] = 0$$

- 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 $x_2 = 10$:

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

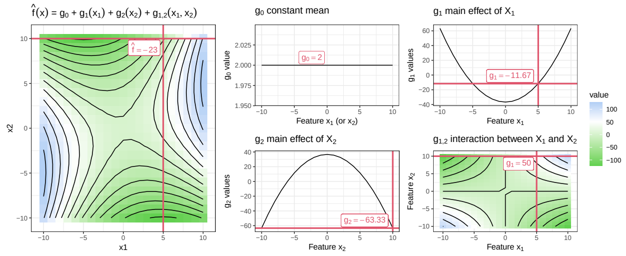
$\Rightarrow \hat{f}(\mathbf{x}) = -23$

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

CONSTRAINTS: EQUIVALENT CHARACTERIZATION

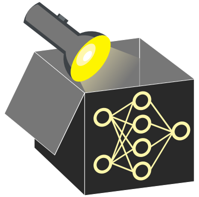
- So far: Definition of standard fANOVA implies vanishing conditions

CONSTRAINTS: EQUIVALENT CHARACTERIZATION

- So far: Definition of standard fANOVA implies vanishing conditions

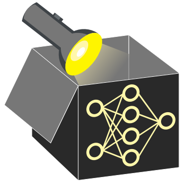
Interpretable Machine Learning – 5 / 5

CONSTRAINTS: EQUIVALENT CHARACTERIZATION



- So far: Definition of standard fANOVA implies vanishing conditions
- Opposite is true as well:
  Features independent $\Longrightarrow$ 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 $\Longrightarrow$ 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 $\Longrightarrow$ 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 $\Longrightarrow$ 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 $\Longrightarrow$ 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 $\Longrightarrow$ 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 $\Longrightarrow$ 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 $\Longrightarrow$ 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