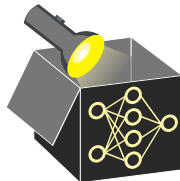


Interpretable Machine Learning

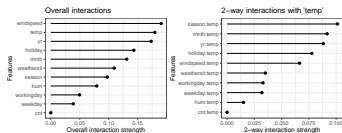
Functional Decompositions

Friedman's H-Statistic



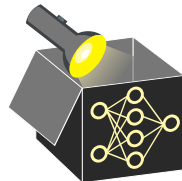
Learning goals

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



2-way interaction:

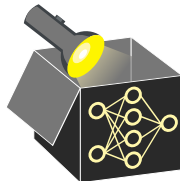
- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0



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:

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



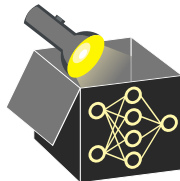
2-way interaction:

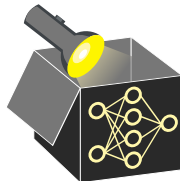
- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

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

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

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





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:

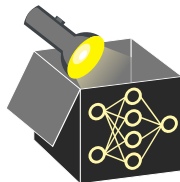
$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_{\emptyset} + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

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

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

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

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



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:

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

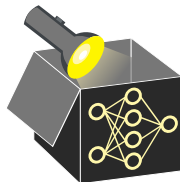
- Here: **Centered PD-functions** $\hat{f}_{S,PD}^c(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_{\emptyset}$

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

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

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

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



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** $\hat{f}_{S,PD}^c(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_{\emptyset}$

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

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

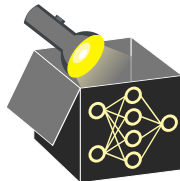
$$\begin{aligned} \hat{f}_{\{jk\},PD}^c(x_j, x_k) &= g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \hat{f}_{\{jk\},PD}^c(x_j, x_k) &= \hat{f}_{j,PD}^c(x_j) + \hat{f}_{k,PD}^c(x_k) \end{aligned}$$

- 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

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:

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



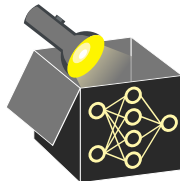
3-way interaction:

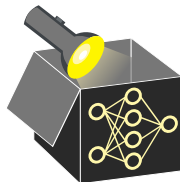
- **Definition:** \hat{f} contains no 3-way interactions between features i, j, k , if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

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

- **Example:**

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





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:

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

- **Example:**

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

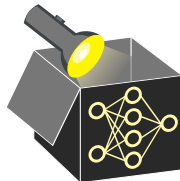
- **Note:** Again using centered PD-functions $\hat{f}_{S,PD}^c$ instead of components g_S
 \leadsto things get complicated, e.g. for 3 features, definition becomes:

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

k -way interaction:

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

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



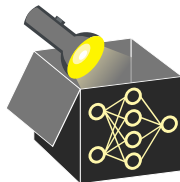
k -way interaction:

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

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

Overall interaction:

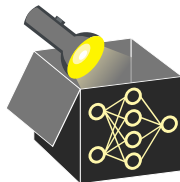
- Question: Does feature j interact with any other feature at all?
- ⇒ H-statistic analogous to 2-way interactions, but for feature sets $S = \{j\}$ and $-S = \{1, \dots, p\} \setminus \{j\}$ instead of two single features:



k -way interaction:

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

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

**Overall interaction:**

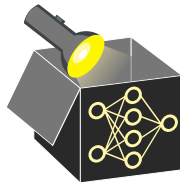
- Question: Does feature j interact with any other feature at all?
- ⇒ H-statistic analogous to 2-way interactions, but for feature sets $S = \{j\}$ and $-S = \{1, \dots, p\} \setminus \{j\}$ instead of two single features:

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

- $-j$ denotes $-S = \{1, \dots, p\} \setminus \{j\}$, 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

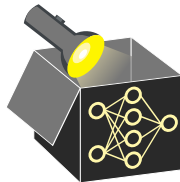
- **Question:** How to measure interaction strength without computing functional decomposition components g_S ?



2-WAY INTERACTION STRENGTH

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

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



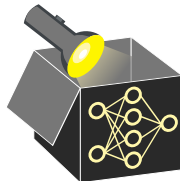
2-WAY INTERACTION STRENGTH

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

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

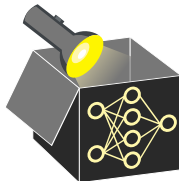
- **H-statistic** for 2-way interaction between feature j and k :

$$\begin{aligned} H_{jk}^2 &= \frac{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) - \hat{f}_{j,PD}^c(X_j) - \hat{f}_{k,PD}^c(X_k) \right]}{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) \right]} \\ &= \frac{\sum_{i=1}^n \left(\hat{f}_{jk,PD}^c(x_j^{(i)}, x_k^{(i)}) - \hat{f}_{j,PD}^c(x_j^{(i)}) - \hat{f}_{k,PD}^c(x_k^{(i)}) \right)^2}{\sum_{i=1}^n \left(\hat{f}_{jk,PD}^c(x_j^{(i)}, x_k^{(i)}) \right)^2} \end{aligned}$$



2-WAY INTERACTION STRENGTH

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



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

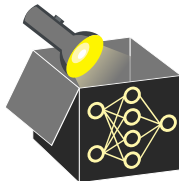
- **H-statistic** for 2-way interaction between feature j and k :

$$\begin{aligned} H_{jk}^2 &= \frac{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) - \hat{f}_{j,PD}^c(X_j) - \hat{f}_{k,PD}^c(X_k) \right]}{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) \right]} \\ &= \frac{\sum_{i=1}^n \left(\hat{f}_{jk,PD}^c(x_j^{(i)}, x_k^{(i)}) - \hat{f}_{j,PD}^c(x_j^{(i)}) - \hat{f}_{k,PD}^c(x_k^{(i)}) \right)^2}{\sum_{i=1}^n \left(\hat{f}_{jk,PD}^c(x_j^{(i)}, x_k^{(i)}) \right)^2} \end{aligned}$$

$\Rightarrow H_{jk}^2$ measures strength of this interaction quantitatively

H_{jk}^2 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

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

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

$$\begin{aligned} H_{jk}^2 &= \frac{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) - \hat{f}_{j,PD}^c(X_j) - \hat{f}_{k,PD}^c(X_k) \right]}{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) \right]} \\ &= \frac{\sum_{i=1}^n \left(\hat{f}_{jk,PD}^c(x_j^{(i)}, x_k^{(i)}) - \hat{f}_{j,PD}^c(x_j^{(i)}) - \hat{f}_{k,PD}^c(x_k^{(i)}) \right)^2}{\sum_{i=1}^n \left(\hat{f}_{jk,PD}^c(x_j^{(i)}, x_k^{(i)}) \right)^2} \end{aligned}$$

$\Rightarrow H_{jk}^2$ measures strength of this interaction quantitatively

H_{jk}^2 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

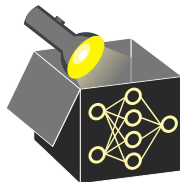
Example

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

$$\hat{f}_{1,PD}^c(x_1) = -2x_1 + 0.5|x_1| + 0.75$$

$$\hat{f}_{2,PD}^c(x_2) = 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05$$

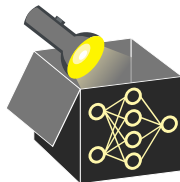
$$\hat{f}_{1,2;PD}^c(x_1, x_2) = 1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e$$



H-STATISTIC: EXAMPLES

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

Example



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

$$\hat{f}_{1,PD}^c(x_1) = -2x_1 + 0.5|x_1| + 0.75$$

$$\hat{f}_{2,PD}^c(x_2) = 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05$$

$$\hat{f}_{1,2;PD}^c(x_1, x_2) = 1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e$$

$$\begin{aligned} \Rightarrow H_{12}^2 &= \frac{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) - \hat{f}_{j,PD}^c(X_j) - \hat{f}_{k,PD}^c(X_k) \right]}{\text{Var} \left[\hat{f}_{jk,PD}^c(X_j, X_k) \right]} \\ &= \frac{\mathbb{E} \left[(|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25)^2 \right]}{\mathbb{E} \left[(1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e)^2 \right]} > 0 \end{aligned}$$

3-WAY INTERACTION STRENGTH

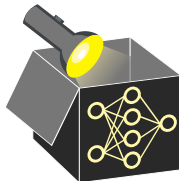
- Same idea as for 2-way, but different formula (see before):

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

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

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

- Analogous for higher order interactions, but more complicated

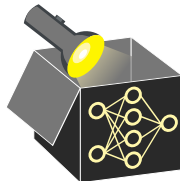


OVERALL INTERACTION STRENGTH

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

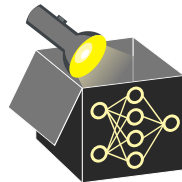
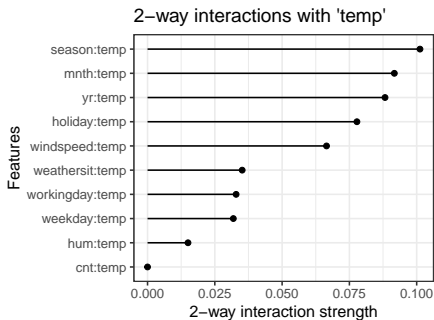
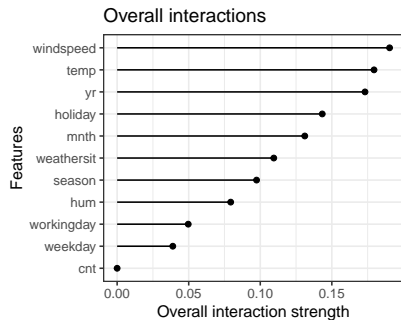
⇒ **H-statistic** analogous to 2-way interaction:

$$\begin{aligned} H_j^2 &= \frac{\text{Var} \left[\hat{f}^c(\mathbf{X}) - \hat{f}_{j,PD}^c(X_j) - \hat{f}_{-j,PD}^c(\mathbf{X}_{-j}) \right]}{\text{Var} \left[\hat{f}^c(\mathbf{X}) \right]} \\ &= \frac{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)}) - \hat{f}_{j,PD}^c(x_j^{(i)}) - \hat{f}_{-j,PD}^c(\mathbf{x}_{-j}^{(i)}) \right)^2}{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)}) \right)^2} \end{aligned}$$



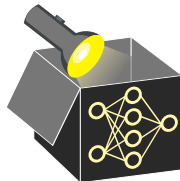
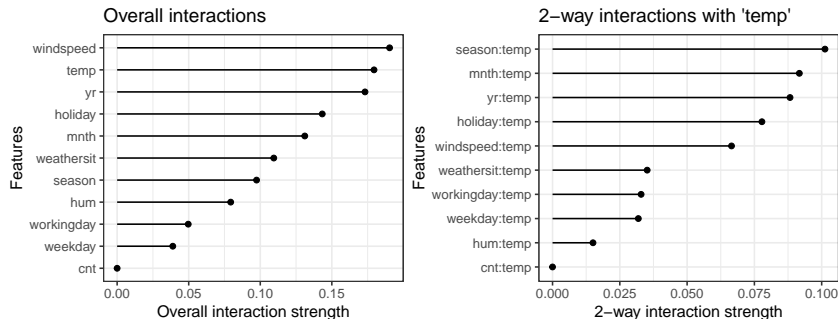
H-STATISTIC: EXAMPLE

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