Achieving interpretable machine learning by functional decomposition of black-box models into explainable predictor effects

Matthias Schmid

joint work with David Köhler, David Rügamer,
Lindsey J. Boyle and Kelly O. Maloney

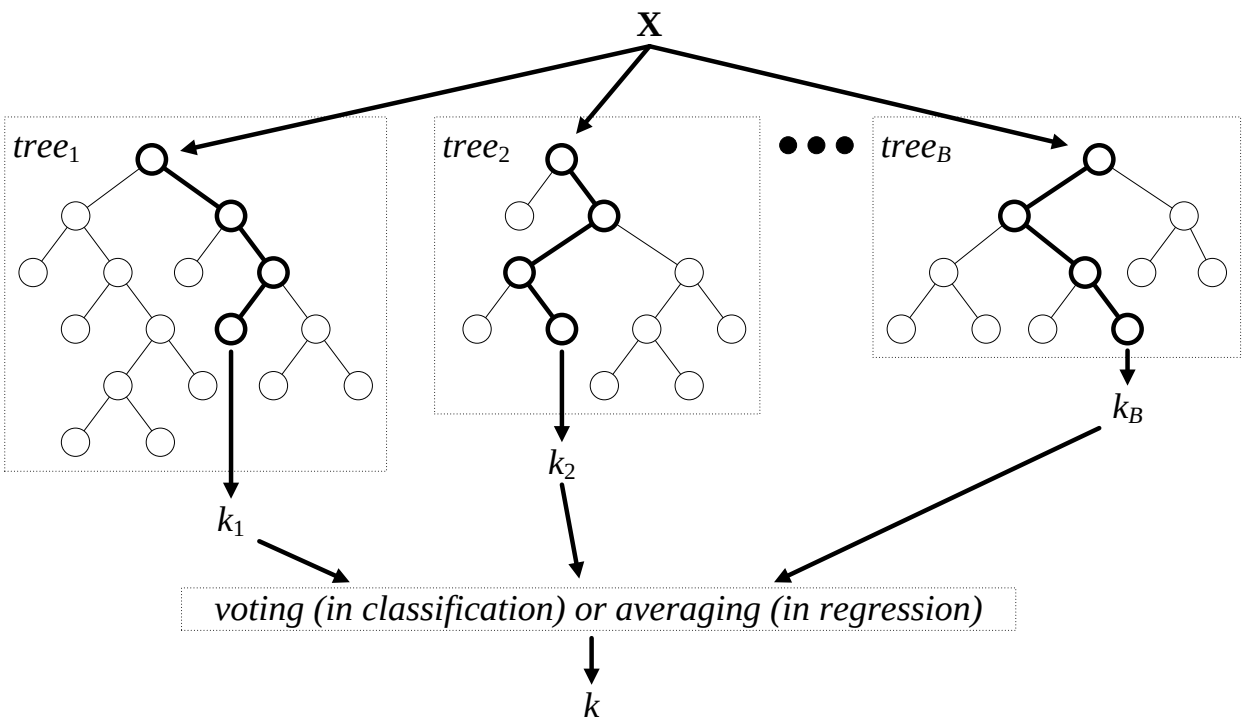
Registry-Based Non-Randomized Studies for Treatment Comparisons
Universitätsmedizin Göttingen, May 26, 2026

Preliminaries

- Supervised machine learning (ML) has seen significant growth in both popularity and importance
- Outcome variable Y, set of features $X = \{X_1, \ldots, X_d\}$
- Predictions of Y are obtained by a prediction function $F(X) \in \mathbb{R}$
- High (prediction) accuracy of supervised ML models often achieved through complex black-box architectures
- Problem: black-box architectures are usually difficult to interpret / explain

Preliminaries

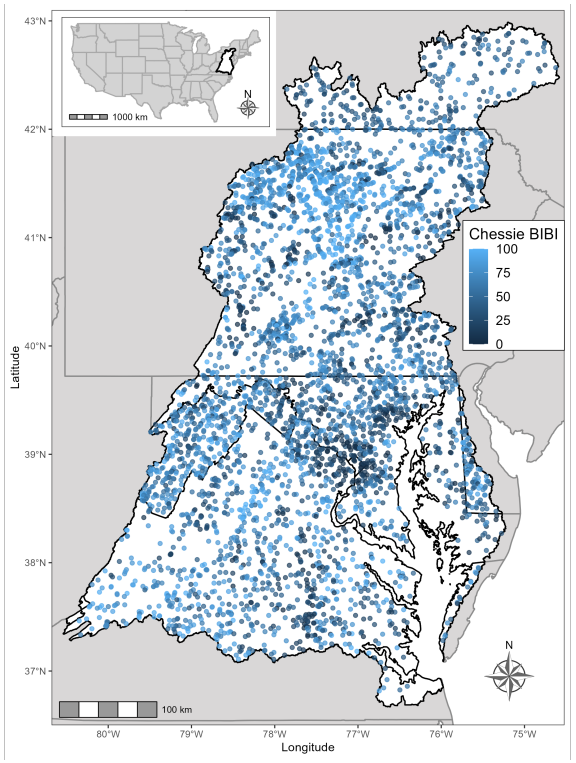
- Example: random forest



Source: Verikas et al. (2016)

Preliminaries

- In Maloney et al. (2022), we fitted a random forest model to analyze stream biological condition in the Chesapeake Bay watershed (USA)



Preliminaries

- Outcome: Basin-wide Index of Biotic Integrity (BIBI)
  - Multi-metric index derived from stream benthic macroinvertebrate samples
  - Measures the biological quality of streams and wadeable rivers
  - Ranges between 0 to 100
- Features: data on climate, land use and watershed characteristics

Preliminaries

- Predictions are intended...
  - ... to estimate stream biological condition at unsurveyed sites
  - ... to inform future management policies (projecting, e.g., changes in land use, climate and watershed characteristics)
- Consequently, estimated/predicted BIBI values are required to be interpretable...
  - ... in terms of relevant features
  - ... in terms of the directions and strengths of the feature effects

Preliminaries

- Need for "opening the black box" has driven research on interpretable machine learning (IML)
- In this talk, we propose a novel approach for the functional decomposition of black-box predictions
- Functional decomposition is a core concept of IML
- Idea: replace the prediction function by a surrogate model consisting of simpler subfunctions (facilitating interpretability)
- Model-agnostic method: can be applied to a broad range of prediction functions, regardless of the type of ML method applied to learn F

Terms and definitions

- Interpretability = "the degree to which a human can understand the cause of a decision" (Miller 2019)
- Explainability = "the internal logic and mechanics that are inside a machine learning system" (Linardatos et al. 2021)
- Use of the terms is ambiguous, and today we will not distinguish between the two
- "Model-based" ("by-design") interpretability
  - Impose an interpretable structure on F during the learning process
- "Post hoc" interpretability (applicable to black-box models)
  - Achieve interpretability by post-processing an already learned prediction model

Terms and definitions

- Examples of post hoc methods
  - Functional decomposition
  - Partial dependence plots (PDP)
  - Accumulated local effects (ALE) plots
  - Shapley values
  - Permutation feature importance
  - ...
- Underlying principle of these methods: measure the variability of F w.r.t. changes in subsets of the features X

Functional decomposition

- Idea: decompose F into a set of simpler ("more interpretable") functions depending on subsets of the features only
- Let $\Upsilon = \{1, \ldots, d\}$ the feature indices and $\mathcal{P}(\Upsilon)$ the power set
- Then F can be decomposed into

$$F(X) = \mu + \sum_{\theta \in \mathcal{P}(\Upsilon):|\theta|=1} f_\theta(X_\theta) + \sum_{\theta \in \mathcal{P}(\Upsilon):|\theta|=2} f_\theta(X_\theta) + \ldots + \sum_{\theta \in \mathcal{P}(\Upsilon):|\theta|=d} f_\theta(X_\theta)\,, \tag{1}$$

- $\mu \in \mathbb{R}$ is an intercept term
- X_θ denotes the subset of features with indices θ, $\theta \in \mathcal{P}(\Upsilon)\backslash\emptyset$
- Example: if $d = 3$ and $\theta = \{1, 3\}$, then $X_\theta = \{X_1, X_3\}$

Functional decomposition

- In IML: main focus is usually on $|\theta| = 1$ (main effects) and $|\theta| = 2$ (two-way interactions)
- Main effects and two-way interactions allow for a simple graphical analysis

  $\Rightarrow$ Line plots, heatmaps, ...
- Functions f_θ with $|\theta| > 2$ ("multivariate feature interactions") are the less interpretable components of F
- In the following we present a novel approach to specify and compute the functions f_θ, given a fixed prediction function F

Conditions on X and f_θ

- Obviously, the functions f_θ are not uniquely defined
- Example: Let $d = 2$, $\mu = 0$ and $F(X_1, X_2) = X_1 + X_1 \cdot X_2$. Then
  - $\{f_1(X_1) = X_1 \quad , f_2(X_2) = 0, f_{12}(X_1, X_2) = X_1 \cdot X_2\}$ and
  - $\{f_1(X_1) = 0.5 \cdot X_1, f_2(X_2) = 0, f_{12}(X_1, X_2) = 0.5 \cdot X_1 + X_1 \cdot X_2\}$

  both satisfy Eq. (1)

$\Rightarrow$ Further assumptions needed to obtain a unique decomposition

Conditions on X and f_θ

- Conditions on X
  - $X_1, \ldots, X_d$ real-valued random variables with bounded support
  - $(X_1, \ldots, X_d)$ defined on a joint probability space P_X
- Conditions on f_θ
  - Each f_θ is square integrable w.r.t. P_X
  - Each f_θ is mean centered, i.e. $\int f_\theta(X_\theta)\, dP_X = 0$
  - f_θ, $\theta \in \mathcal{P}(\Upsilon)\backslash\emptyset$, are linearly independent
- Further definitions
  - $\sigma_\theta^2 = \int f_\theta^2(X_\theta)\, dP_X$ – variance of f_θ
  - $\sigma_{\theta\theta'} = \int f_\theta(X_\theta)\, f_{\theta'}(X_{\theta'})\, dP_X$ – covariance of f_θ and $f_{\theta'}$, $\theta, \theta' \in \mathcal{P}(\Upsilon)\backslash\emptyset$

Principles for the decomposition

- Main requirement: The summands in Eq. (1) should be well separated
- In particular, higher-order effects (with large $|\theta|$) should not contain any components of lower-order effects (with small $|\theta|$)
- Related concepts:
  - Purity criterion (Molnar 2022): predictive information explained by a main effect is not contained in the higher-order effects that include the corresponding feature
  - Optimality criterion (Hooker 2007): lower-order functions should capture as much functional behavior as possible

Generalized functional ANOVA

- To implement optimality, Hooker (2007) proposed a decomposition termed generalized functional ANOVA
- With this approach, the functions in Eq. (1) are required to be hierarchically orthogonal, satisfying

$$\forall \theta \in \mathcal{P}(\Upsilon) \backslash \emptyset \quad \forall \theta' \subseteq \theta : \\ \sigma_{\theta\theta'} = \int f_\theta(X_\theta) f_{\theta'}(X_{\theta'}) dP_X = 0 \tag{2}$$

$\Rightarrow$ For any given θ', the effect $f_{\theta'}(X_{\theta'})$ is orthogonal to all higher-order effects $f_\theta(X_\theta)$ with $X_\theta \supseteq X_{\theta'}$

Generalized functional ANOVA

- Generalized functional ANOVA has been acknowledged as a key concept for making ML models interpretable
- Problem: computational and numerical issues associated with the calculation of the feature effects f_θ
- In the following, we present...
  - ... the concept of stacked orthogonality (alternative approach to implement purity/optimality)
  - ... a user-friendly algorithm to estimate the functions f_θ
  - ... a coefficient to measure the degree of explainability of a black-box model

Stacked orthogonality

- We require the functions f_θ to meet the stacked orthogonality constraints

$$\forall k \in \Upsilon : \int \Big(\sum_{\substack{\theta \in \mathcal{P}(\Upsilon): \\ |\theta| = k}} f_\theta(X_\theta) \Big) \Big(\sum_{\substack{\theta' \in \mathcal{P}(\Upsilon): \\ |\theta'| < k}} f_{\theta'}(X_{\theta'}) \Big) dP_X = 0 \,, \tag{3}$$

  where $k \in \Upsilon$ denotes the effect level (= interaction order)

- For each k, the sum of the level-k effects is required to be uncorrelated with the sum of all lower-level effects

- Level-wise implementation of purity/optimality

Degree of interpretability

- Convenient feature of stacked orthogonality: The variance of F can be decomposed in a level-wise fashion
- Idea: Define the fraction of σ_F^2 explained by the k-th level by

$$I_k = \frac{\int \left(\sum_{\theta' \in \mathcal{P}(\Upsilon):|\theta'|=k} f_{\theta'}(X_{\theta'}) \right)^2 dP_X}{\sigma_F^2} \tag{4}$$

- Degree of interpretability $= I_1 + I_2$

Estimation by neural additive models with post-hoc orthogonalization

- Aim: develop a user-friendly algorithm to compute the functions f_θ (satisfying the stacked orthogonality constraints)
- We propose the following three-step procedure:

Step 1 Sample training data

Step 2 Fit a neural additive model (NAM) to the training data

Step 3 Apply post-hoc orthogonalization to the NAM fit

Estimation by neural additive models with post-hoc orthogonalization

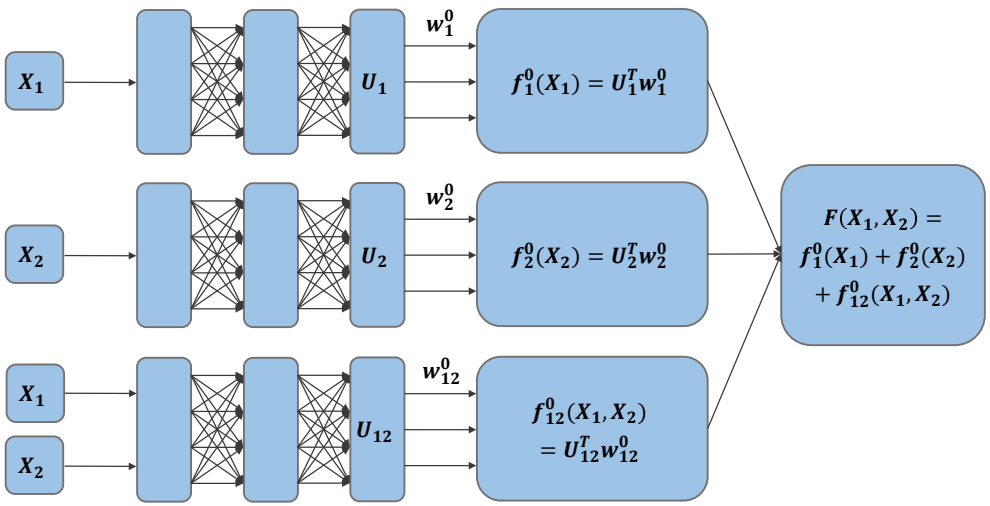
- Step 1
  - Generate a sample of n data points $\{F_i, X_{i1}, \ldots, X_{id}\}_{i=1,\ldots,n}$
  - X_{ij}, $j \in \Upsilon$, and $F_i = F(\{X_{i1}, \ldots, X_{id}\})$ denote the j-th feature value and the i-th value of the prediction function, respectively
- Step 2
  - Train a neural additive model (NAM) of the form

$$F_i = \sum_{\theta \in \mathcal{P}(\Upsilon):|\theta|=1} f_\theta^0(X_{i\theta}) + \sum_{\theta \in \mathcal{P}(\Upsilon):|\theta|=2} f_\theta^0(X_{i\theta}) + \ldots + \sum_{\theta \in \mathcal{P}(\Upsilon):|\theta|=d} f_\theta^0(X_{i\theta}), \quad i = 1, \ldots, n, \quad (5)$$

where $X_{i\theta}$ are the sample values of X_θ

Estimation by neural additive models with post-hoc orthogonalization

- Illustration of a NAM (Agarwal et al. 2021)



- Model fitting is performed using backpropagation, with each function f_θ^0 represented by an artificial neural (sub)network (ANN)
- ANNs can approximate general classes of functions arbitrarily well

Estimation by neural additive models with post-hoc orthogonalization

- Step 3
  - Apply post-hoc orthogonalization to the NAM fit
  - Extension of an approach by Rügamer (2023)
- Post-hoc orthogonalization is done in an iterative fashion, starting at the highest level and proceeding down to the first level
- Strategy: at each level k, ...

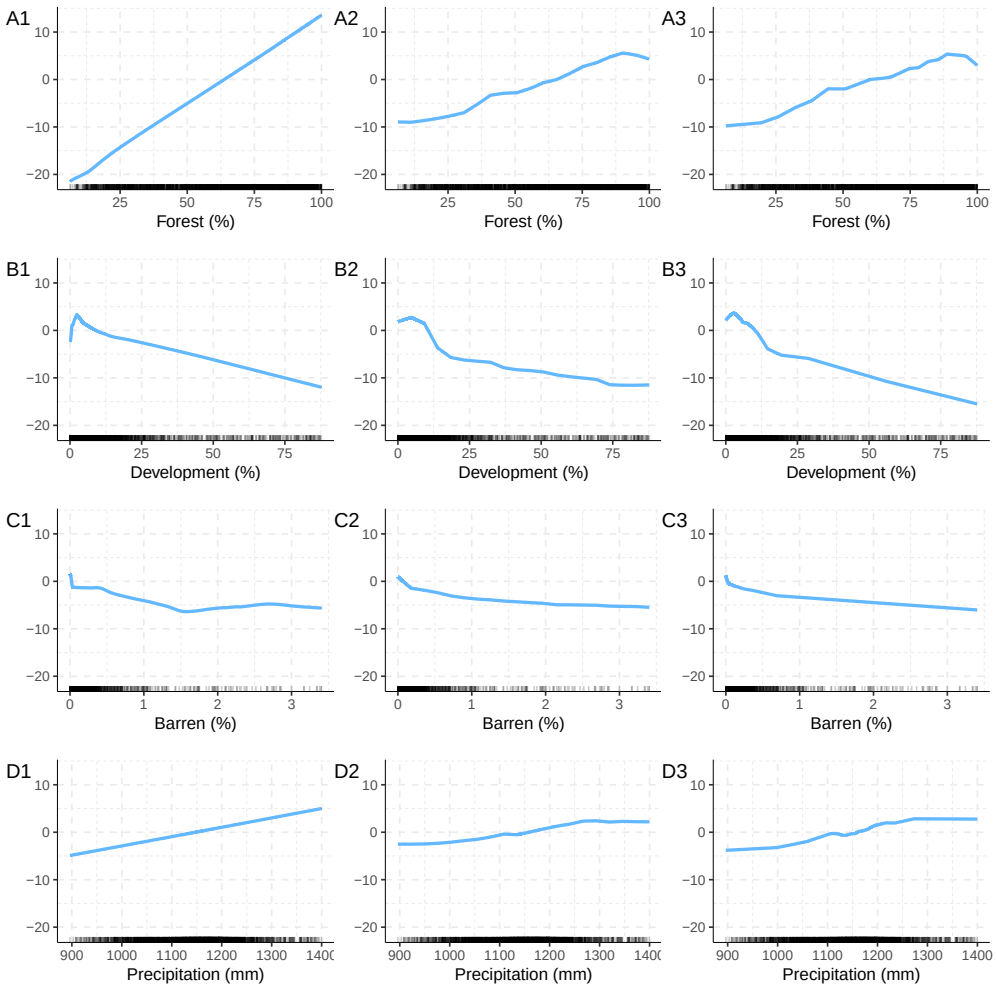
  ... project the sum of the k-way interactions onto the column space spanned by lower-order interactions ($< k$),

  ... update k-way interactions by functions orthogonal to this space,

  ... update lower-order interactions by adding projections,

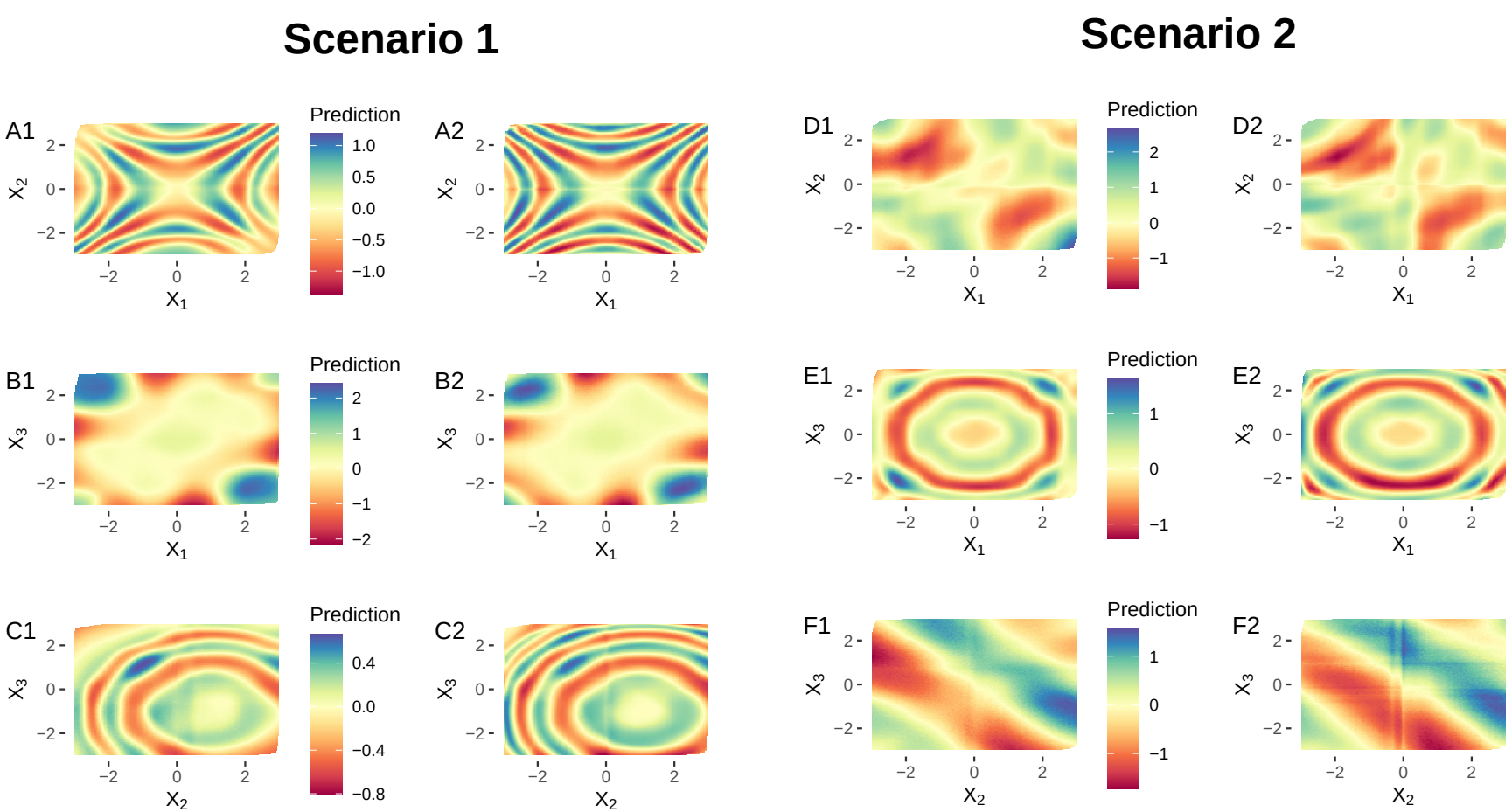
  ... leave higher-order interactions ($> k$) unchanged

Illustration – random forest model by Maloney et al. (2022)



$l_1 = 0.908$

Illustration – synthetic data, two-way interactions



Remarks on NAMs and post-hoc orthogonalization

- For NAM fitting, we propose an ensemble strategy with different weight initializations
  - Idea: better approximation of highly non-linear effects by averaged NAM fits
- Important: we do not use NAMs for supervised learning
  - Procedure does not involve Y
  - Instead, the predicted values F_i are used as outcome of the NAM
- We do not want to avoid overfitting
  - F_i should be approximated as closely as possible
  - There is no residual error term in Eq. (5)

Remarks on NAMs and post-hoc orthogonalization

- Post-hoc orthogonalization does not require re-fitting the NAM ($\rightarrow$ efficient)
- Implemented in R package ONAM
  https://cran.r-project.org/web/packages/ONAM
- For regularity assumptions and details on the implementation, see Köhler et al. (2025)

Further remarks

- The proposed method is designed to explain the *inner* workings of a black-box model
- It can not be used to evaluate the features' ability to predict Y

$\Rightarrow$ The effects obtained from our method will only have a useful interpretation if F is good for prediction

- Stacked orthogonality approach can be adapted to model sets of "effects of interest"
  - Let $\Theta \subset \mathcal{P}(\Upsilon)\backslash\emptyset$ represent the effects of interest
  - Then $\mathcal{P}(\Upsilon)\backslash(\Theta \cup \emptyset)$ can be absorbed into the last summand f^0_Υ
  - NAM fitting and post-hoc orthogonalization can be applied as before
  - This strategy is necessary when d is large

References

Agarwal R, Melnick L, Frosst N, Zhang X, Lengerich B, Caruana R, Hinton G, Neural additive models: Interpretable machine learning with neural nets. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) (Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds) 34 (Advances in Neural Information Processing Systems, 2021), 4699-4711.

Hooker G, Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. Journal of Computational and Graphical Statistics, 16, 709-732, (2007).

Linardatos P, Papastefanopoulos V, Kotsiantis S, Explainable AI: A review of machine learning interpretability methods. Entropy, 23:18, (2021).

Maloney KO, Buchanan C, Jepsen RD, Krause KP, Cashman MJ, Gressler BP, Young JA, Schmid M, Explainable machine learning improves interpretability in the predictive modeling of biological stream conditions in the Chesapeake Bay Watershed, USA. Journal of Environmental Management, 322, 116068, (2022).

Miller T, Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38, (2019).

References

Molnar C, Interpretable Machine Learning – A Guide for Making Black Box Models Explainable, 2nd ed. (Independently published, 2022).

Rügamer D, A new PHO-rmula for improved performance of semi-structured networks. In Proceedings of the 40th International Conference on Machine Learning (eds, A. Krause et al.) 202 (Proceedings of Machine Learning Research, 2023), 29291-29305.

Verikas A, Vaiciukynas E, Gelzinis A, Parker J, Olsson MC, Electromyographic patterns during golf awing: Activation sequence profiling and prediction of shot effectiveness. Sensors 16:592, (2016).

This talk:

Köhler D, Rügamer D, Boyle LJ, Maloney KO, Schmid M (2025). Achieving interpretable machine learning by functional decomposition of black-box models into explainable predictor effects. Npj Artificial Intelligence 1:34.