ICML
International Conference
On Machine Learning

Exact Functional ANOVA Decomposition for Categorical Inputs Models

Baptiste Ferrere · Nicolas Bousquet · Fabrice Gamboa
Jean-Michel Loubes · Joseph Muré

ICML 2026 – Seoul, South Korea

Authors

B. Ferrere[*]
EDF R&D, SINCLAIR
Univ. Toulouse

N. Bousquet[†]
EDF R&D, SINCLAIR
Univ. Sorbonne

F. Gamboa[†]
Univ. Toulouse
ANITI
Univ. Medellín

JM. Loubes[†]
Univ. Toulouse
ANITI
Inria REGALIA

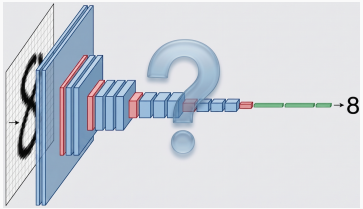
J. Muré[†]
EDF R&D, SINCLAIR

[*] Corresponding author [†] Equal supervision

baptiste.ferrere@edf.fr ▪ bapfr.github.io ▪ BapFr ▪ Google Scholar

eXplainable Artificial Intelligence (XAI)

Millions of non-linear operations



LLM hallucinations

User: How many `r`'s in *strawberry*?

Model: There are **two** `r`'s in strawberry.

!

Feature importance

Misclassification

The many additive explanations

Additive explanation. Decompose the prediction $f(\mathbf{x})$ for a given input $\mathbf{x} \coloneqq (x_1, \dots, x_d)$ in term of feature contributions:

$$f(\mathbf{x}) = \text{Contrib.}(x_1) + \text{Contrib.}(x_2) + \cdots + \text{Contrib.}(x_d)$$

$\Rightarrow$ we analyse the impact of each feature on the output.

Some well known methods:

- **Sampling-based:** LIME (Ribeiro'16), SHAP (Lundberg & Lee'17)
- **Gradient-based:** DeepLIFT (Shrikumar'17), LRP (Bach'15), IG (Sundararajan'17)
- **Functional ANOVA** (Hoeffding'48, Stone'94, Hooker'07) $\leftarrow$ *Our focus*

Functional ANOVA: $f(\mathbf{x}) = \sum_{A \subseteq [d]} f_A(\mathbf{x}_A) = \text{cst} + \sum_i f_i(\mathbf{x}_i) + \sum_{i<j} f_{i,j}(\mathbf{x}_i, \mathbf{x}_j) + \dots$

Two cases for Functional ANOVA

Independent case

Hoeffding'48 – Stone'94

Möbius transform:

$$f_A(\mathbf{X}_A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \, \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_B]$$

- Closed form ✓
- Pairwise orthogonality ✓
- Variance decomposition ✓

Dependent case

Hooker 2007, Chastaing et al. 2012, Il Idrissi et al. 2025

$$f_A(\mathbf{X}_A) = \; ?$$

no explicit construction

- Existence & uniqueness only
- Hierarchical orthogonality required
- Costly sampling-based approximations

Our Contribution

Analytical Closed-Form Solution

We derive a **closed-form formulation** to compute $f_A(\mathbf{X}_A)$ for categorical data under any arbitrary distribution p.

Dependency-Aware: Explicitly accounts for complex feature dependencies.

Computational Efficiency: Enables real-time interpretability indices.

Links with existing literature

We connect many objects such as **Classical Orthogonal ANOVA**, **Generalized ANOVA**, **Boolean Fourier Analysis** and **Shapley Values**.

The Central Object

Definition: Inverse Likelihood Mechanism

For each $(A, \mathbf{z}) \in \mathcal{I}$, we define the basis function $\phi_A^{(\mathbf{z})}$ as:

$$\phi_A^{(\mathbf{z})}(\mathbf{x}) \;:=\; 1/p_A(\mathbf{x}_A) * \prod_{i \in A} \left(\mathbf{1}_{x_i = z_i} - \mathbf{1}_{x_i = N_i - 1}\right)$$

Structural Notations

- **Categorical Input:** $X_i \in \{0, \ldots, N_i - 1\}$
- **Index set:** $\mathcal{I}$ contains the pairs $(A, \mathbf{z})$ s.t. $A \subseteq [d]$ and $z_i < N_i - 1 \; \forall i \in A$.
- **Denominator:** p_A marginal pmf of $\mathbf{X}_A$

Key Properties

- **Classical ANOVA:** Recovers orthogonality in independent setting
- **Fourier Link:** Extends the Walsh-Hadamard basis

Main Result: Functional Representation

Theorem: Universal ANOVA Expansion

Any function $f \in L^2(p)$ always admits an expansion of the form:

$$f(\mathbf{X}) = \sum_{(A,\mathbf{z})\in\mathcal{I}} c_A^{(\mathbf{z})} \cdot \phi_A^{(\mathbf{z})}(\mathbf{X})$$

General Case

On **full support**, the expansion is the unique solution to the generalized functional ANOVA variational problem (Hooker '07).

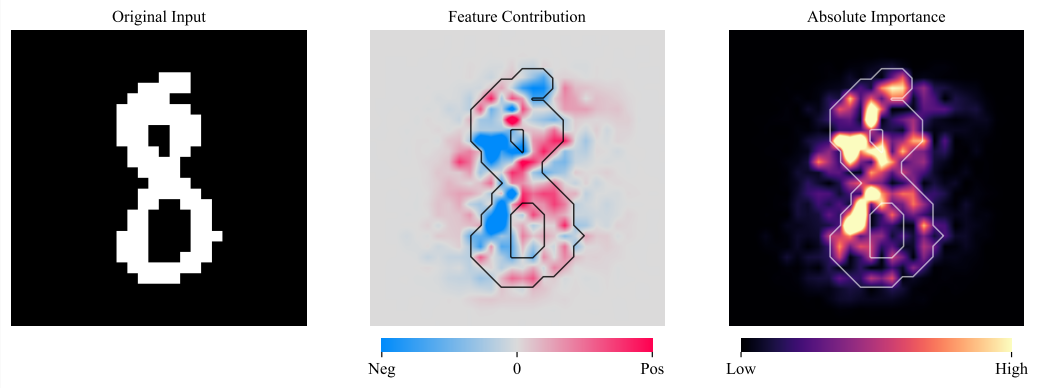
Independence

Basis becomes pairwise orthogonal, recovering the classical Möbius formula.

Boolean Limit

Recovers Boolean Fourier analysis on the hypercube (O'Donnell'14).

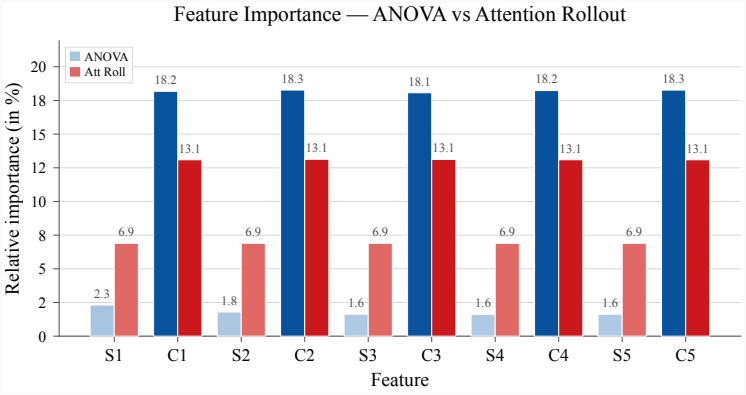
Visual Example: Opening the Black Box on MNIST



Red *– pixels* ***boosting*** *the probability to classify the image as a 3.*
Blue *–* ***suppressing*** *it (evidence maintaining the '8' classification).*

60k samples explained once

Tabular Transformer Example : Feature Attribution on Poker Hands



Domain Logic Recovery:
ANOVA (Ours) correctly identifies **Ranks (C)** as the primary features, while **Suits (S)** are marginal.

Attention Rollout:
Over-attributes importance to suits, failing to capture rules.

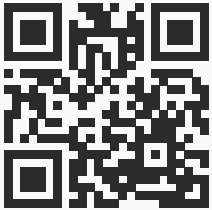
Thank you!!!

Paper

arXiv:2603.02673

Code

github.com/BapFr



Website

bapfr.github.io

baptiste.ferrere@edf.fr