GENERALIZED FUNCTIONAL ANOVA: A COMPLETE THEORETICAL FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

The functional ANOVA provides a fundamental representation of square-integrable multivariate functions into main effects and higher-order interactions. For independent inputs, the components belong to mutually orthogonal Hilbert subspaces and admit an explicit representation. For dependent inputs, however, the components are only hierarchically orthogonal: although existence and uniqueness results are available, the Hilbert subspaces underlying the generalized decomposition have remained implicit. We resolve this representation problem for continuous inputs supported on a bounded hyperrectangle whose joint density is bounded above and away from zero. We introduce a distribution-adapted family of functions and prove that it forms a Riesz basis of the L^2 space, thereby guaranteeing a unique, stable, and unconditionally convergent representation. We then show that, for every coalition of variables, the corresponding block of this basis exactly characterizes the Hilbert subspace containing the functional ANOVA component. Our construction recovers the classical orthogonal decomposition under input independence. As a direct consequence, computing the generalized functional ANOVA reduces to estimating coefficients in an explicit, distribution-adapted basis. Finally, as a *proof of concept*, we introduce an elementary, fast and model-agnostic estimator based on our theoretical results. Experiments on synthetic and real-world datasets illustrate its connections with established tabular explanation methods and show that low order components often capture most of the signal in the model output.

1 INTRODUCTION

Notations. Let $\mathbb{N}_+$ denote the set of positive integers and let take $d \in \mathbb{N}_+$. We denote by $[d]$ the set $\{1, \dots, d\}$. We also denote by $\mathbf{X}$ (in bold) the real valued random vector $(X_1, \dots, X_d)$ whose distribution is denoted $P_{\mathbf{X}}$. For any subset S of $[d]$, we define $\mathbf{X}_S := (X_i)_{i \in S}$. Finally, we denote by $L^2(P_{\mathbf{X}})$ the Hilbert space of $P_{\mathbf{X}}-$square integrable functions.

The functional ANOVA decomposition provides a fundamental representation of multivariate functions. It expresses any $f \in L^2(P_{\mathbf{X}})$ as follows:

$$f(\mathbf{X}) = f_\emptyset + \sum_{i=1}^{d} f_i(X_i) + \sum_{i<j} f_{i,j}(X_i, X_j) + \cdots + f_{[d]}(\mathbf{X}) = \sum_{S \subseteq [d]} f_S(\mathbf{X}_S). \tag{1}$$

Each component f_S depends only on the variables indexed by S and represents the contribution associated with this subset after removing all lower-order effects. When the inputs are mutually independent, these components belong to mutually orthogonal Hilbert subspaces and admit an explicit expression through the Möbius inversion of conditional expectations (Hoeffding, 1948). The functional ANOVA therefore provides both a decomposition of individual functions and an orthogonal decomposition of the underlying $L^2(P_{\mathbf{X}})$ space.

This classical construction fundamentally relies on input independence. When the inputs are dependent, the component subspaces are no longer mutually orthogonal, and the usual conditional-expectation formula does not isolate pure interaction effects. The identification principle now commonly referred to as *hierarchical orthogonality* already appears in the functional ANOVA framework of Stone (1994) and was subsequently made explicit by Huang (1998). More precisely, for every $S \subseteq [d]$, the S-order component is required to be orthogonal to every function belonging to a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

space indexed by a strict subset $T \subsetneq S$. Under suitable bounded-density assumptions, these works obtain an essentially unique theoretical ANOVA decomposition. Hooker (2007) subsequently recast this construction for dependent inputs as a constrained functional optimization problem and call it the *generalized* functional ANOVA. Later works investigated its existence and uniqueness under different dependence assumptions (Chastaing et al., 2012; Idrissi et al., 2025). Nevertheless, the geometry of this generalized decomposition remains implicit: the Hilbert subspaces containing its components have not been explicitly characterized through a basis of $L^2(P_{\mathbf{X}})$ well suited to the distribution $P_{\mathbf{X}}$.

This representation problem is directly relevant to statistical learning theory. A full functional ANOVA decomposes an arbitrary multivariate function into components of every order, whereas additive learning imposes a structural restriction by retaining only a small collection of these components. Generalized additive models retain first-order effects (Hastie, 2017), while GA^{2}Ms additionally include selected pairwise interactions (Lou et al., 2013). Related ideas underlie spline-based functional ANOVA (Stone, 1994), component-selection methods such as sparse additive models (Ravikumar et al., 2009; Raskutti et al., 2012; Haris et al., 2022). By exploiting sparsity and low-order interactions, these models provide flexible nonparametric representations while mitigating the curse of dimensionality.

Additive and low-order interaction structures have recently attracted renewed attention in machine learning. Bhattacharya et al. (2024) established minimax-optimal rates for suitably constructed deep neural networks over nonparametric interaction models of fixed order. At the architectural level, Neural Additive Models (Agarwal et al., 2021), NODE-GAM (Chang et al., 2021), and Neural Basis Models (Radenovic et al., 2022) combine neural-network estimation with additive functional representations. These developments illustrate the continuing importance of identifying and estimating low-dimensional components of a multivariate function. They also emphasize the need for a precise characterization of the spaces to which these components belong, particularly when the inputs are not independent.

Functional decompositions have also become increasingly important for explainable AI. Inherently interpretable models expose low-order functional components by construction. In parallel, several post-hoc explanation methods are closely connected to the functional ANOVA. Under input independence, ANOVA components can be interpreted as Harsanyi dividends (Harsanyi, 1963) allowing to compute Shapley attributions (Shapley, 1953; Lundberg and Lee, 2017). This results from redistributing each interaction component among its participating variables (Owen, 2014; Owen and Prieur, 2017). Functional ANOVA has also been used to obtain *pure* interaction effects in additive models and to explain tree ensembles (Lengerich et al., 2020). Very recently, (Bénard, 2025) develops a method from data, to estimate the functional ANOVA of a trained tree ensemble, highlights strong connections with TreeSHAP (Lundberg et al., 2018) and SHAP-based functional decompositions of tree ensembles (Herren and Hahn, 2022; Bordt and von Luxburg, 2023; Muschalik et al., 2024a;b). Finally, Ferrere et al. (2026) show that, for categorical inputs, the generalized functional ANOVA admits an explicit finite-dimensional representation based on inverse likelihood weighting. This result provides a finite-dimensional precursor to the construction studied here. Extending it to continuous inputs is, however, not obvious: the resulting representation becomes infinite-dimensional, and completeness or linear independence alone is insufficient to warrant existence, uniqueness, unconditional convergence, and norm stability. These properties require establishing the substantially stronger Riesz-basis property for the dictionary of functions that gives a representation of the functional decomposition.

Thus, rather than constituting an isolated explanation method, functional ANOVA provides a common mathematical object connecting inherently additive predictors and post-hoc additive explanations.

More precisely, in this work, we focus on random vectors $\mathbf{X} \in [-1, 1]^d$ whose joint density p is bounded above and away from zero, a standard regularity condition (see for example Stone (1977)) and largely used in the analysis of generalized functional ANOVA and nonparametric function estimation (Stone, 1994; Györfi et al., 2002; Chastaing et al., 2012; Bénard, 2025). We introduce a distribution-adapted family obtained by weighting a Hilbert basis by inverse marginal densities and prove that it forms a Riesz basis of $L^2(P_{\mathbf{X}})$. We use this basis to characterize explicitly the Hilbert subspace containing every generalized functional ANOVA component and derive its unique and stable expansion. Finally, we introduce a simple and fast model-agnostic explainer as a proof of concept and use it to illustrate the connections between generalized functional ANOVA and established tabular explanation methods. We detail our contribution in the next paragraph.

Contribution.

1. We introduce in Definition 3.1 a novel distribution-adapted family of functions. We prove in Theorem 3.2 that this family forms a *Riesz basis* of $L^2(P_{\mathbf{X}})$, thereby guaranteeing a unique, stable, and unconditionally convergent representation of every function in this Hilbert space.
2. We characterize the Hilbert subspaces underlying the generalized functional ANOVA. Specifically, Theorem 3.4 shows that, for every $S \subseteq [d]$, the component f_S belongs to a fully characterized subspace, denoted by $\mathcal{H}_S$, and admits a unique expansion in the corresponding block of our Riesz basis. To the best of our knowledge, this is the first explicit characterization of these subspaces for dependent continuous inputs. As a direct consequence, computing the generalized functional ANOVA reduces to a standard statistical learning problem: estimating coefficients in a Riesz basis.
3. We demonstrate in Corollary 3.5 that, under input independence, our construction exactly recovers the classical orthogonal decomposition. In particular, the subspaces $\mathcal{H}_S$ become mutually orthogonal and the resulting components coincide with the classical functional ANOVA components.
4. As a *proof of concept* for our theoretical results, we propose a simple full data-driven algorithm, without knowing the architecture of the model f or the distribution $P_{\mathbf{X}}$. Through experiments on synthetic and real-world datasets, we illustrate how the decomposition behaves in practice and highlight its close empirical connections with well-established tabular XAI methods. We further observe that, across the considered datasets and model classes, restricting the decomposition to first- and second-order components is often sufficient to capture most of the variance in the model output.

2 Preliminaries

We will focus on the expansion (1), where each component $f_S(\mathbf{X}_S)$ depends only on the subset of variables $\mathbf{X}_S$. The constant term $f_\emptyset$ captures the overall mean effect, the functions f_i represent the marginal (or main) effects, and higher-order terms f_S with $|S| \geq 2$ encode the pure interaction effects among the variables indexed by S. In this paper, we assume that $\mathbf{X}$ has a joint continuous density p, *i.e.* $dP_{\mathbf{X}}(\mathbf{x})/d\mathbf{x} = p(\mathbf{x})$.

2.1 Generalized Functional ANOVA

For all $S \subseteq [d]$, let L^2_S be the Hilbert space of square integrable functions of $\mathbf{X}_S$. One can define the components of (1) as the solution of the following optimization problem (Hooker, 2007):

$$\{f_S\}_{S\subseteq[d]} \coloneqq \underset{\{g_S \in L^2_S\}_{S\subseteq[d]}}{\operatorname{argmin}} \int_{\mathbb{R}^d} \left(\sum_{S\subseteq[d]} g_S(\mathbf{x}_S) - f(\mathbf{x}) \right)^2 p(\mathbf{x})\, d\mathbf{x}, \tag{2}$$

subject to the *hierarchical orthogonality constraint*:

$$\forall T \subsetneq S \subseteq [d],\ \forall\, g_T \in L^2_T,\ \int_{\mathbb{R}^d} f_S(\mathbf{x}_S)\, g_T(\mathbf{x}_T)\, p(\mathbf{x})\, d\mathbf{x} = 0. \tag{3}$$

Conceptually, decomposition (1) is sought under the constraint that information is non-redundant: for any subset $T \subseteq [d]$, the contribution of f_S for $T \subsetneq S \subseteq [d]$ must capture only the variance that cannot be explained by any function of a strict subset of its variables $\mathbf{X}_T$. Roughly speaking, each component f_S is *purely an $|S|$-th order effect*, free of any lower-order information.

Under mild non-degeneracy conditions on the joint distribution (Chastaing et al., 2012; Idrissi et al., 2025), the optimization problem (2) under the constraint (3) admits a unique solution, and the resulting decomposition generally allocates less variance to higher-order components. Throughout this work, we assume that the support of $\mathbf{X}$ is a hyperrectangle and impose the following regularity condition on the density (Stone, 1977; 1994).

Assumption 2.1. Without loss of generality, we assume that $\mathbf{X}$ is supported on $[-1, 1]^d$ and that its density p is bounded above and away from zero: there exist constants $0 < c_1 < c_2$ such that:

$$\forall\, \mathbf{x} \in [-1, 1]^d,\ c_1 \leq p(\mathbf{x}) \leq c_2 < \infty. \tag{4}$$

Theorem 2.2 (Hooker (2007)). *For any function f member of $L^2(P_{\mathbf{X}})$, there exists a unique collection of functions $\{f_S\}_{S\subseteq[d]}$ such that $f(\mathbf{X}) = \sum_{S\subseteq[d]} f_S(\mathbf{X}_S)$ under the hierarchical orthogonality constraint (3).*

2.2 Standard Functional ANOVA

In this short subsection, we recall the basics results under independence, i.e., when the underlying density p is product. In particular, we have the following theorem due to Hoeffding (1948) and Stone (1994).

Theorem 2.3 (Standard functional ANOVA). *Under mutual independence, the components $f_S(\mathbf{X}_S)$ are explicit and are given by the Möbius transform Rota (1964) of the conditional expectations:*

$$\forall S \subseteq [d],\ f_S(\mathbf{X}_S) = \sum_{T\subseteq S} (-1)^{|S|-|T|} \mathbb{E}\left[f(\mathbf{X}) \mid \mathbf{X}_T\right]. \tag{5}$$

Moreover, the constraint (3) implies the mutual orthogonality of the decomposition:

$$\forall S, T \subseteq [d],\ S \neq T \implies \mathbb{E}[f_S(\mathbf{X}_S)\, f_T(\mathbf{X}_T)] = 0, \tag{6}$$

recovering the well-known orthogonal variance decomposition.

A strong connection with Shapley values. Functional ANOVA is closely connected to Shapley values through Harsanyi dividends (Harsanyi, 1963). For a fixed input $\mathbf{x}$, the components $f_S(\mathbf{x}_S)$ are the dividends of the cooperative game with value function $v_{\mathbf{x}}(S) := \sum_{T\subseteq S} f_T(\mathbf{x}_T)$. Distributing each dividend equally among its participating features yields the *fANOVA Shapley values* (Owen, 2014; Owen and Prieur, 2017): $\mathrm{shap}_i^{\mathrm{ANOVA}}(\mathbf{x}) := \sum_{S\ni i} f_S(\mathbf{x}_S)/|S|$. Under mutual independence of the input features, these attributions coincide with SHAP values for the conditional-expectation game (Lundberg and Lee, 2017). In this paper, we further investigate the connection empirically by comparing estimated ANOVA main effects with SHAP-based explanations. The visual comparisons in the main body of the paper and in Appendix F; and the global feature-importance rank correlations in Appendix E, provide complementary qualitative and quantitative evidence of strong overall agreement across the considered datasets and model classes.

3 Theoretical Characterization

In this section, we establish our main theoretical results. We first define a novel family of functions that generalizes the inverse likelihood mechanism of Ferrere et al. (2026) to continuous inputs. We then prove that this family forms a Riesz basis of $L^2(P_{\mathbf{X}})$. A Riesz basis (see for example Brézis (2011)) is the image of an orthonormal basis under a bounded invertible operator; it preserves uniqueness, stability, and unconditional convergence while allowing for non-orthogonality. Finally, we show that this basis is adapted to the generalized functional ANOVA and provides an explicit solution to the optimization problem formulated by Hooker (2007).

3.1 The Distribution-Adapted Riesz Basis

Throughout this section (and the rest of the paper), let first denote by p_S the marginal density of $\mathbf{X}_S$ for every non-empty set $S \subseteq [d]$. For each $m \in \mathbb{N}_+$, let ϕ_m denote the normalized Legendre polynomial of degree m, whose definition and properties are recalled in Appendix A.

Definition 3.1. First, let $\xi_\emptyset := 1$. Then, for every set $S \subseteq [d]$, denote $\boldsymbol{m}_S := (m_j)_{j\in S} \in \mathbb{N}_+^{|S|}$. For $\mathbf{x} \in [-1,1]^d$ we set:

$$\xi_S^{(\boldsymbol{m}_S)}(\mathbf{x}) := \frac{1}{\sqrt{2^{d-|S|}}} \cdot \frac{\prod_{j\in S} \phi_{m_j}(x_j)}{p_S(\mathbf{x}_S)} \tag{7}$$

We denote by Ξ the resulting collection of functions. This construction provides a continuous counterpart to the inverse likelihood mechanism of Ferrere et al. (2026): the sign functions used in the categorical setting are replaced by normalized Legendre polynomials, while inverse probability-mass weighting is replaced by inverse marginal-density weighting. In Appendix B, we show that there exists an *invertible* and *bounded* operator $\mathcal{T}$ such that Ξ is the direct image of an orthonormal basis by this operator $\mathcal{T}$. As a consequence, we have the next representation theorem:

Theorem 3.2 (Riesz Basis)**.** *The family of functions Ξ introduced in Definition 3.1 forms a Riesz basis of $L^2(P_{\mathbf{X}})$.*

As a direct consequence, there exist *Riesz constants* $C_{\min}, C_{\max} > 0$ (which are independent of summation indices S and $\boldsymbol{m}_S$) such that, for every sequence $a \in \ell^2$, the following spectral inequality is satisfied:

$$C_{\min} \|\boldsymbol{a}\|_{\ell_2}^2 \leq \left\| \sum_{S \subseteq [d]} \sum_{\boldsymbol{m}_S \in \mathbb{N}_+^{|S|}} a_S^{(\boldsymbol{m}_S)} \, \xi_S^{(\boldsymbol{m}_S)} \right\|_{L^2(P_{\mathbf{X}})}^2 \leq C_{\max} \|\boldsymbol{a}\|_{\ell_2}^2 . \tag{8}$$

In the very particular case where the basis is an orthonormal (or also said an Hilbert) basis, we have $C_{\min} = C_{\max} = 1$.

3.2 The Representation Theorem

Before obtaining the main theorem, we introduce the following Lemma that is the desired consequence of the inverse likelihood mechanism.

Lemma 3.3 (Hierarchical Orthogonality)**.** *The family of functions Ξ introduced in Definition 3.1 satisfies hierarchical orthogonality in $L^2(P_{\mathbf{X}})$. More formally, we have:*

$$\forall T \subsetneq S \subseteq [d], \forall \boldsymbol{m}_S \in \mathbb{N}_+^{|S|}, \forall \boldsymbol{n}_T \in \mathbb{N}_+^{|T|}, \quad \mathbb{E}\left[\xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}) \cdot \xi_T^{(\boldsymbol{n}_T)}(\mathbf{X})\right] = 0. \tag{9}$$

Let Assumption 2.1 hold, thanks to this previous technical result and to the Riesz basis property, we have all the ingredients to state the following theorem, which is the central contribution of the paper.

Theorem 3.4 (Main Theorem)**.** *For any $f \in L^2(P_{\mathbf{X}})$, there exists a unique sequence of coefficients $\boldsymbol{\theta}^\star := \left(\theta_S^{(\boldsymbol{m}_S)}\right)_{S \subseteq [d],\, \boldsymbol{m}_S \in \mathbb{N}_+^{|S|}}$ such that the components satisfying (2) and (3) express as follows:*

$$f_S = \sum_{\boldsymbol{m}_S \in \mathbb{N}_+^{|S|}} \theta_S^{(\boldsymbol{m}_S)} \, \xi_S^{(\boldsymbol{m}_S)}, \qquad \forall\, S \subseteq [d]. \tag{10}$$

By denoting the sequence of Hilbert spaces $(\mathcal{H}_S)_{S \subseteq [d]}$ associated to the ANOVA decomposition of $L^2(P_{\mathbf{X}})$ (*i.e.* the hierarchical Hilbert spaces that will contain all the functions f_S, see Huang (1998); Idrissi et al. (2025)), this theorem is strictly equivalent to the following characterization:

$$\mathcal{H}_S = \overline{\text{span}} \left\{ \xi_S^{(\boldsymbol{m}_S)} \;\middle|\; \boldsymbol{m}_S \in \mathbb{N}_+^{|S|} \right\}. \tag{11}$$

3.3 The Independent Setting

As discussed before, under feature independence, the constraint (3) strengthens to full mutual orthogonality between components. In what follows, we show that our basis recovers naturally this property.

Corollary 3.5 (Orthogonality)**.** *Let Assumption 2.1 hold. Let also assume that $P_{\mathbf{X}}$ is a product measure over $[-1, 1]^d$, then the elements of Ξ are mutually orthogonal set by set, i.e.*

$$\forall T \neq S \subseteq [d], \forall \boldsymbol{m}_S \in \mathbb{N}_+^{|S|}, \forall \boldsymbol{n}_T \in \mathbb{N}_+^{|T|}, \quad \mathbb{E}\left[\xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}) \cdot \xi_T^{(\boldsymbol{n}_T)}(\mathbf{X})\right] = 0. \tag{12}$$

As a direct consequence of this corollary, the resulting decomposition is mutually orthogonal, which means that:

$$\forall S, T \subseteq [d],\ S \neq T \implies \mathbb{E}[f_S(\mathbf{X}_S)\, f_T(\mathbf{X}_T)] = 0, \tag{13}$$

so our formulation entirely recovers the *standard* setting. This result is also proved in Appendix B.

Remark 3.6 (Uniform Setting). In the uniform independent setting, the blocks of Ξ coincide, up to nonzero subset-dependent scaling factors, with the tensorized Legendre basis. Consequently, they recover the same mutually orthogonal decomposition.

4 Estimating the Decomposition

In this section, we assume that an i.i.d. sample $\{\mathbf{X}^{(i)}\}_{i=1}^n$ is at hand together with the evaluations $\{f(\mathbf{X}^{(i)})\}_{i=1}^n$. Here, f is treated as a *black-box* model: we can query it pointwise but have no access to its internal architecture (e.g., a trained neural network or tree ensemble).

4.1 A Nonparametric Estimation Problem

Since the basis Ξ is available in closed form, estimating the functional ANOVA decomposition reduces to estimating the coefficients:

$$\boldsymbol{\theta}^\star = \left(\left(\theta_S^{(\boldsymbol{m}_S)} \right)_{\boldsymbol{m}_S \in \mathbb{N}_+^{|S|}} \right)_{S \subseteq [d]}. \tag{14}$$

A natural estimator is obtained by minimizing the general least-squares risk given by:

$$\mathcal{L}_n(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^n \left(f(\mathbf{X}^{(i)}) - \sum_{S \subseteq [d]} \sum_{\boldsymbol{m}_S \in \mathbb{N}_+^{|S|}} \beta_S^{(\boldsymbol{m}_S)} \cdot \xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}^{(i)}) \right)^2 + \mathrm{pen}(\boldsymbol{\beta}). \tag{15}$$

The Riesz-basis representation turns the computation of the generalized functional ANOVA into the estimation of an infinite-dimensional sequence of coefficients. Abstracting away from the non-orthogonal geometry of the basis, this is a classical problem of nonparametric series estimation: the target function is represented by an infinite coefficient sequence, approximated within a sequence of finite-dimensional spaces, and the corresponding coefficients are estimated by empirical least squares (Newey, 1997; Györfi et al., 2002; Chen, 2007; Tsybakov, 2009; Cohen et al., 2013).

However, developing a complete nonparametric theory for this formulation—including the choice of truncation or regularization, the approximation properties of the resulting sieve spaces, the effect of estimating the distribution-adapted basis, and the derivation of convergence or minimax rates—raises substantial statistical questions that require a separate treatment and lie beyond the scope of this paper. We therefore do not pursue the asymptotic analysis of the estimator here. Instead, in the next subsection, we introduce a deliberately simple finite-dimensional estimator under explicit structural assumptions. This estimator is intended as a *proof of concept* to illustrate the practical and computational implications of our theoretical representation.

4.2 A Truncated Series Estimator

To obtain a finite-dimensional problem, we truncate the Riesz expansion jointly with respect to the interaction order and the coordinatewise polynomial degree. This construction follows standard truncation strategies from the polynomial chaos expansion literature (Xiu and Karniadakis, 2002; Tsybakov, 2009; Blatman and Sudret, 2011). The restriction on the interaction order is motivated by the *sparsity-of-effects principle* (Montgomery, 2017), according to which the response is often primarily driven by main effects and low-order interactions, and by the *reluctance principle* (Yu et al., 2019), which favors lower-order effects whenever they provide comparable explanatory power.

Definition 4.1 (Truncation set)**.** Let $K \in [d]$ and $M \in \mathbb{N}_+$. We define the truncation set of maximum interaction order K and maximum coordinatewise degree M by

$$\mathcal{I}_K^M := \left\{ (S, \boldsymbol{m}_S) \,\middle|\, S \subseteq [d],\ |S| \le K,\ \boldsymbol{m}_S \in \{1, \dots, M\}^{|S|} \right\}, \tag{16}$$

where the empty set is associated with the unique empty multi-index.

Remark 4.2 (Dimension of the truncated space). The truncation set contains

$$\mathcal{N}(d, K, M) := \sum_{k=0}^K \binom{d}{k} M^k \le \underbrace{M^K}_{\text{exponential complexity}} * \underbrace{\left(\frac{\mathrm{e}\, d}{K} \right)^K}_{\text{Sauer-Shelah complexity}} \tag{17}$$

elements. Consequently, its dimension is polynomial in d and M when K is fixed, but grows combinatorially with the maximum interaction order.

This truncation is consistent with the low-order structures underlying GAMs, GA^{2}Ms, and fixed-order interaction models (Nori et al., 2019; Agarwal et al., 2021; Bhattacharya et al., 2024; Bénard, 2025). It does not, however, enforce sparsity among components of the same order, since every subset S satisfying $|S| \leq K$ is retained. Such effect selection could be incorporated through an additional penalty on the coefficient blocks. Here, we deliberately retain the simplest construction and empirically examine whether main and pair effects capture most of the model output.

The basis functions depend on the marginal densities of the input distribution, which are unknown in practice. We estimate these densities using truncated orthogonal-series estimators constructed from normalized Legendre polynomials and denote by $\widehat{\xi}_S^{(\boldsymbol{m}_S)}$ the resulting plug-in basis functions. Restricting the expansion to $\mathcal{I}_K^M$ then yields the finite-dimensional estimator

$$\widehat{\boldsymbol{\theta}}_n^{K,M} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^{\mathcal{N}(d,K,M)}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{X}^{(i)}) - \sum_{(S,\boldsymbol{m}_S)\in\mathcal{I}_K^M} \beta_S^{(\boldsymbol{m}_S)}\widehat{\xi}_S^{(\boldsymbol{m}_S)}(\mathbf{X}^{(i)})\right)^2. \tag{18}$$

When the empirical design matrix has full column rank, the minimizer is unique. Otherwise, we define $\widehat{\boldsymbol{\theta}}_n^{K,M}$ as the minimum-Euclidean-norm least-squares solution. In Appendix C, Figure 3 summarizes the overall procedure and we also provide the complete estimation pipeline.

Quantifying how the truncation error and the plug-in estimation of the input densities jointly affect the estimation of the generalized functional ANOVA belongs to the broader statistical analysis identified above is not pursued in this paper.

5 Experiments

Consistent with the theoretical scope of this paper, the experiments are intended as a *proof of concept* rather than as a comprehensive statistical evaluation of the proposed estimator. They investigate two main questions. (i) Reconstruction fidelity, *i.e.* how good the truncated and estimated decomposition reconstruct the model ; (ii) Agreement with additive explainers, *i.e.* how these components align with standard additive explanation methods. We first consider an analytically tractable example for which the generalized functional ANOVA components are available in closed form. We display on Figure 1 theoretical univariate functions and visually compare them with the classic explanation methods. We then conduct experiments on six real-world and classical datasets given in the Table 1.

ID	Name	n	d
BS	*Bike Sharing* (Fanaee-T and Gama, 2014)	17 379	8
CH	*California Housing* (Pace and Barry, 1997)	20 640	8
CI	*Census Income* (Kohavi et al., 1996)	48 842	14
EG	*Electrical Grid* (Arzamasov et al., 2018)	10 000	12
PI	*Pima Indian Diabetes* (Smith et al., 1988)	768	8
SC	*Superconductivity* (Hamidieh, 2018)	21 263	81

Table 1: Datasets characteristics.

For both analytical and real-world dataset settings, we train two black-box predictors—a neural network and an XGBoost model—using fixed model specifications detailed in Appendix D. We apply our decomposition to both predictors and evaluate the reconstruction fidelity of our low order truncated ANOVA. We additionally visually compare the resulting components with several established explanation methods, using each method with the model class for which it is designed. For neural networks, we consider KernelSHAP (Lundberg and Lee, 2017), a widely used model-agnostic attribution method, and DeepSHAP (Shrikumar et al., 2017; Lundberg and Lee, 2017), its model-specific counterpart. For tree ensembles, we consider TreeSHAP (Lundberg et al., 2018), the standard SHAP implementation for tree-based models, and TreeHFD (Bénard, 2025), which directly estimates a functional ANOVA decomposition of trained tree ensembles. Finally, we train Neural Additive Models (NAMs, Agarwal et al. (2021)) and Explainable Boosting Machines (EBMs, Nori et al. (2019)), apply our decomposition to the resulting predictors. We also evaluate the reconstruction fidelity and visually compare the recovered main effects with the components natively produced by these architectures.

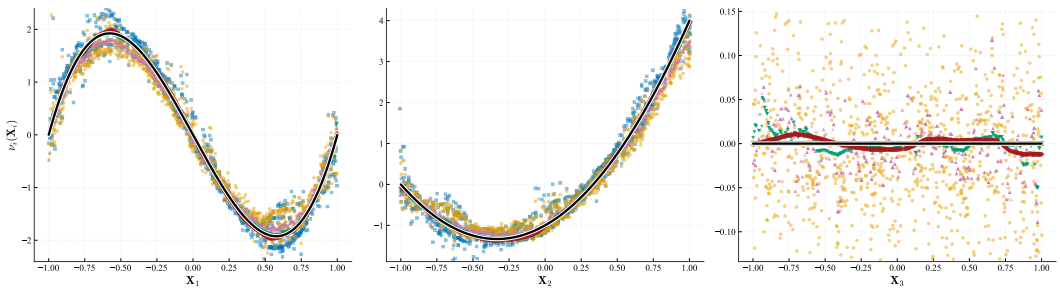


Figure 1: Estimated main effects in the analytical setting for X_1, X_2, and the irrelevant variable X_3. All methods (KernelSHAP, TreeSHAP, TreeHFD, NAM, EBM) closely track the **theoretical ANOVA component (in black)** on X_1 and X_2, and correctly assign a near-zero contribution to X_3.

Analytical Case. We consider a random vector $\mathbf{X} \in [-1, 1]^3$ of following density

$$\forall \mathbf{x} \in [-1, 1]^3, \quad p_\rho(\mathbf{x}) := \frac{1}{8}\left(1 + \rho(x_1x_2 + x_1x_3 + x_2x_3)\right), \tag{19}$$

with $\rho \in (-1/3, 1)$ so that p_ρ satisfies Assumption 2.1. The target function is defined as

$$f(\mathbf{X}) := f_1(X_1) + f_2(X_2) + f_{1,2}(X_1, X_2), \tag{20}$$

whose components are given in closed form in Appendix D.4. This setting has two convenient features: the marginal and bivariate densities admit explicit expressions and the variable X_3 is irrelevant to f. This allows us to check if different methods correctly assign it a zero contribution. We display these theoretical components and the corresponding contributions given by the post-hoc explanation methods and interpretable models in Fig 1.

Real-world datasets. The datasets characteristics are summarized in Table 1. We also present the fidelity reconstruction in Table 2. We illustrate in Figure 2 an example of visual agreement (all other figures are deferred to the Appendix).

Table 2: Reconstruction fidelity of truncated functional ANOVA.

Dataset	Model	Acc.	K	Rec.	Ortho.
BS (17 379, 8)	XGB	0.86	2	0.92	9.49e−2
	MLP	0.85	2	0.91	8.83e−2
	EBM	0.83	2	0.96	9.52e−2
	NAM	–	–	–	–
CH (20 640, 8)	XGB	0.84	2	0.86	6.35e−2
	MLP	0.81	1	0.88	0.00
	EBM	0.83	2	0.89	4.04e−2
	NAM	0.76	1	0.94	0.00
CI (48 842, 14)	XGB	0.88	1	0.91	0.00
	MLP	0.85	1	0.94	0.00
	EBM	0.88	1	0.90	0.00
	NAM	–	–	–	–

Dataset	Model	Acc.	K	Rec.	Ortho.
EG (10 000, 12)	XGB	0.93	2	0.89	7.41e−3
	MLP	0.98	2	0.95	6.61e−3
	EBM	0.94	2	0.99	5.58e−3
	NAM	0.88	1	1.00	0.00
PI (768, 8)	XGB	0.74	2	0.85	9.56e−2
	MLP	0.73	2	0.99	0.00
	EBM	0.76	2	0.97	0.00
	NAM	0.73	1	1.00	0.00
SC (21 263, 81)	XGB	0.93	1	0.86	0.00
	MLP	0.91	1	0.89	0.00
	EBM	0.90	1	0.88	0.00
	NAM	–	–	–	–

Briefly, **Acc.** (for *Accuracy*) quantifies the predictive performance of the trained machine learning model; $K \in \{1, 2\}$ is the truncation level of our decomposition; **Rec.** (for *Reconstruction*) is the reconstruction R^2 of our estimator, *i.e.* the part of the variance from the trained machine learning model that is explained by our framework; and **Ortho.** (for *Hierarchical Orthogonality*) is a cosine-based quantity measuring the *worst* deviation from hierarchical orthogonality where 0 means perfect hierarchical orthogonality. The more detailed definition of all performance metrics are given in Appendix D.

Results. Across the four considered model families, truncations involving only main effects and pairwise interactions ($K \leq 2$) achieve a reconstruction R^2 above 0.8 on every dataset. This indicates

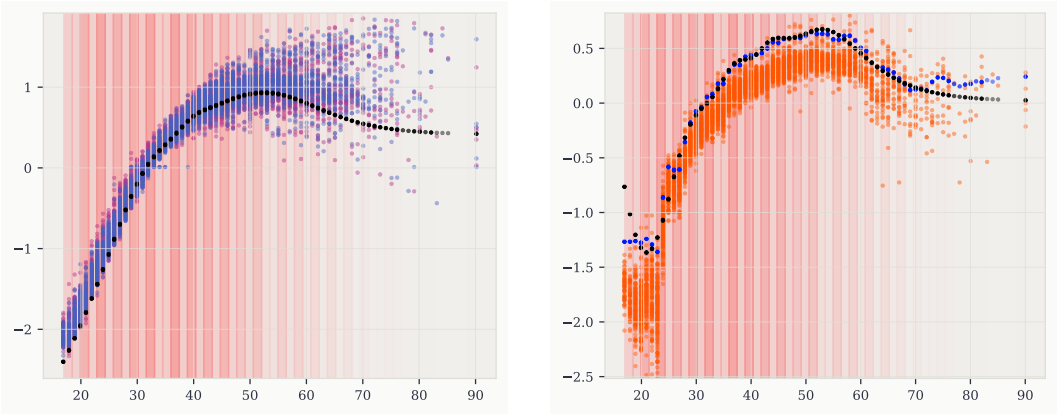


Figure 2: Estimated first-order effect of `Age` on the *Census Income* dataset.
Left: our functional ANOVA estimate (black) compared with KernelSHAP and DeepSHAP attributions for a trained MLP. **Right:** our estimate (black) compared with the TreeHFD main effect and TreeSHAP attributions for a trained XGBoost model. The intensity of the red background represents the empirical marginal density of `Age`, with darker regions indicating higher data concentration. Our component follows the central trend of the SHAP attributions and closely agrees with TreeHFD, particularly in high-density regions.

that low-order components capture most of the variability in the fitted-model predictions in our experiments, without excluding the presence of higher-order effects. Despite the deliberately simple plug-in estimation of the marginal densities, our empirical diagnostics show only small deviations from the hierarchical-orthogonality conditions. Finally, the runtimes reported in Table 6 demonstrate that the resulting explanations can be computed efficiently while requiring only black-box evaluations of the model. Appendix D provides the complete experimental material, including the full collection of figures, numerical tables and detailed computation-time table.

6 Discussion

Conclusion. In this work, we completely and explicitly characterized the generalized functional ANOVA components for continuous inputs on a bounded domain whose joint density is bounded above and away from zero. This representation turns the computation of the decomposition from an implicit constrained optimization problem into the estimation of coefficients in an explicit basis, thereby connecting it to classical nonparametric series estimation. As a proof of concept, we introduced a simple and model-agnostic estimator. Our experiments show that its low-order components reconstruct a large proportion of the fitted-model predictions and reveal close empirical connections with established additive explanation methods.

Limitations & Future Work. The main limitation of this work is the absence of a statistical analysis of the estimation of the generalized functional ANOVA. Two distinct settings must be considered. First, when the input density is known, the resulting basis is oblique rather than orthogonal, requiring an adaptation of the standard nonparametric estimation literature which could notably draw on the compressed sensing literature. Second, when the density is unknown, appropriate regularity, or smoothness, classes must be introduced for both the target function and the density. One must then determine how the smoothness of the density affects the overall convergence rate when the target function has a different smoothness. A natural next step is therefore to define distribution-adapted global regularity classes, such as Sobolev-type ellipsoids expressed through the coefficients of our Riesz basis, and to construct estimators attaining minimax-optimal rates over these classes, potentially through suitably penalized least squares. Such a development would provide the statistical counterpart to our representation theorem and constitute a substantial contribution to the field of statistical learning theory.

AI USE STATEMENT

In this work, we used generative AI tools for editing and English-language proofreading, as well as for assisting with literature review. In particular, we requested the LLM feedback provided by Paper Assitant Tool (PAT) available on OpenReview for the conference. The main ideas, problem formulation, contributions, theorem statements, and proofs were conceived, formulated, and written by the authors. Generative AI tools were used only subsequently to review, edit, and improve this material. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The results presented in the main body of the paper are demonstrated in the dedicated appendix section (Appendix B). Our proposed estimator is presented in Section 4. The corresponding algorithm and detailed procedure to compute it are provided in Appendix C. Detailed descriptions of the datasets, models, and metrics used in our experiments are given in Appendix D.

REFERENCES

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34:4699–4711.

Arzamasov, V., Böhm, K., and Jochem, P. (2018). Towards concise models of grid stability. In *2018 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm)*, pages 1–6. IEEE.

Bénard, C. (2025). Tree Ensemble Explainability through the Hoeffding Functional Decomposition and TreeHFD Algorithm. *Advances in Neural Information Processing Systems*.

Bhattacharya, S., Fan, J., and Mukherjee, D. (2024). Deep neural networks for nonparametric interaction models with diverging dimension. *The Annals of Statistics*, 52(6):2738–2766.

Blatman, G. and Sudret, B. (2011). Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of computational Physics*, 230(6):2345–2367.

Bordt, S. and von Luxburg, U. (2023). From Shapley values to generalized additive models and back. In *International Conference on Artificial Intelligence and Statistics*, pages 709–745. PMLR.

Brézis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer.

Cencov, N. N. (1962). Estimation of an unknown distribution density from observations. *Soviet Math.*, 3:1559–1566.

Chang, C.-H., Caruana, R., and Goldenberg, A. (2021). NODE-GAM: Neural generalized additive model for interpretable deep learning. *arXiv:2106.01613*.

Chastaing, G., Gamboa, F., and Prieur, C. (2012). Generalized Hoeffding-Sobol Decomposition for Dependent Variables – Application to Sensitivity Analysis. *Electronic Journal of Statistics*, 6:2420–2448.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

Cohen, A., Davenport, M. A., and Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Foundations of computational mathematics*, 13(5):819–834.

Courant, R. and Hilbert, D. (2024). *Methods of mathematical physics, volume 2*. John Wiley & Sons.

Dua, D., Graff, C., et al. (2017). Uci machine learning repository, 2017. *URL http://archive. ics. uci. edu/ml*, 7(1):62.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2).

Fanaee-T, H. and Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2):113–127.

Ferrere, B., Bousquet, N., Gamboa, F., Loubes, J.-M., and Muré, J. (2026). Exact functional ANOVA decomposition for categorical inputs models. In *Forty-third International Conference on Machine Learning*.

Ghanem, R. G. and Spanos, P. D. (2003). *Stochastic finite elements: a spectral approach*. Courier Corporation.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer.

Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354.

Haris, A., Simon, N., and Shojaie, A. (2022). Generalized sparse additive models. *Journal of machine learning research*, 23(70):1–56.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Harsanyi, J. C. (1963). A Simplified Bargaining Model for the $n-$Person Cooperative Game. *International Economic Review*, 4(2):194–220.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Herren, A. and Hahn, P. R. (2022). Statistical aspects of SHAP: Functional ANOVA for model interpretation. *arXiv:2208.09970*.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Hooker, G. (2007). Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *The annals of statistics*, 26(1):242–272.

Idrissi, M. I., Bousquet, N., Gamboa, F., Iooss, B., and Loubes, J.-M. (2025). Hoeffding decomposition of functions of random dependent variables. *Journal of Multivariate Analysis*, 208:105444.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kohavi, R. et al. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207.

Lengerich, B., Tan, S., Chang, C.-H., Hooker, G., and Caruana, R. (2020). Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *International Conference on Artificial Intelligence and Statistics*, pages 2402–2412. PMLR.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 623–631.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv:1802.03888*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons.

Muschalik, M., Baniecki, H., Fumagalli, F., Kolpaczki, P., Hammer, B., and Hüllermeier, E. (2024a). shapiq: Shapley interactions for machine learning. *Advances in Neural Information Processing Systems*, 37:130324–130357.

Muschalik, M., Fumagalli, F., Hammer, B., and Hüllermeier, E. (2024b). Beyond TreeSHAP: Efficient computation of any-order shapley interactions for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14388–14396.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168.

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Owen, A. B. (2014). Sobol'indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251.

Owen, A. B. and Prieur, C. (2017). On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002.

Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radenovic, F., Dubey, A., and Mahajan, D. (2022). Neural basis models for interpretability. *Advances in Neural Information Processing Systems*, 35:8414–8426.

Raskutti, G., J Wainwright, M., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of machine learning research*, 13(2).

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030.

Reed, M. and Simon, B. (1980). Volume 1: Functional analysis. In *Methods of Modern Mathematical Physics*. Elsevier.

Rota, G.-C. (1964). On the foundations of combinatorial theory: I. theory of Möbius functions. In *Classic Papers in Combinatorics*, pages 332–360. Springer.

Shapley, L. S. (1953). *A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261.

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, pages 595–620.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, pages 118–171.

Szegő, G. (1939). *Orthogonal polynomials*, volume 23. American Mathematical Soc.

Tretter, C. (2008). *Spectral theory of block operator matrices and applications*. World Scientific.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Xiu, D. and Karniadakis, G. E. (2002). The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644.

Yu, G., Bien, J., and Tibshirani, R. (2019). Reluctant interaction modeling. *arXiv:1907.08414*.

APPENDIX

A Legendre Polynomials **15**

B Proofs **16**

B.1 Proof of Theorem 3.2 16

B.2 Proof of Lemma 3.3 19

B.3 Proof of Theorem 3.4 19

B.4 Proof of Corollary 3.5 20

C Estimation **21**

D Experiments details **23**

D.1 Resources 23

D.2 Machine learning models 23

D.3 Explainers 24

D.4 Analytical Case 25

D.5 Datasets 26

D.6 Hyperparameters 26

D.7 Performance metrics 27

D.8 Additional Analytical Case 29

E Connection with SHAP-based estimators **30**

F Additional Figures **33**

F.1 Bike Sharing Dataset 33

F.2 California Housing Dataset 34

F.3 Census Income Dataset 35

F.4 Electrical Grid Dataset 36

F.5 Diabetes Dataset 38

A LEGENDRE POLYNOMIALS

In this section, we rely on standard results about Legendre polynomials (see for example Szegő (1939)).

Definition A.1 (Legendre polynomial)**.** For any non-negative integer $m \in \mathbb{N}$, we denote by $\mathbf{P}_m$ the Legendre polynomial of degree m. This family is uniquely defined by the three-term recurrence relation:

$$\begin{cases} \mathbf{P}_0(X) = 1, \\ \mathbf{P}_1(X) = X, \\ (m+1)\,\mathbf{P}_{m+1}(X) = (2m+1)\,X\,\mathbf{P}_m(X) - m\,\mathbf{P}_{m-1}(X), \quad \forall\, m \geq 1. \end{cases} \tag{21}$$

Proposition A.2. *The Legendre polynomials satisfy the following identity:*

$$\forall n, m \in \mathbb{N}, \quad \int_{-1}^{1} \mathbf{P}_n(x)\mathbf{P}_m(x)dx = \frac{2}{2n+1}\mathbf{1}_{\{n=m\}}. \tag{22}$$

Definition A.3. For any integer $m \in \mathbb{N}$, we denote by ϕ_m the *normalized* Legendre polynomial of degree m defined as:

$$\phi_m \coloneqq \sqrt{\frac{2m+1}{2}}\mathbf{P}_m. \tag{23}$$

Theorem A.4. *Reed and Simon (1980); Ghanem and Spanos (2003); Courant and Hilbert (2024) For any multi-index $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, we define the following tensorized product of normalized Legendre polynomials:*

$$\Phi_{\boldsymbol{\alpha}}(X_1, \ldots, X_d) \coloneqq \prod_{j=1}^{d} \phi_{\alpha_j}(X_j). \tag{24}$$

The collection $(\Phi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ forms a Hilbert basis (i.e. an orthonormal basis) of $L^2([-1,1]^d)$ equipped with the Lebesgue measure λ.

A direct consequence of the previous theorem is an *ANOVA-type indexing* of the Hilbert basis of $L^2(\lambda)$, in which each basis element depends only on a prescribed subset of variables. Recall that $\mathbb{N}_+$ denotes the set of positive integers. For any $S \subseteq [d]$ and any $\boldsymbol{m}_S \coloneqq (m_j)_{j\in S} \in \mathbb{N}_+^{|S|}$, set

$$\psi_S^{(\boldsymbol{m}_S)} \coloneqq \frac{1}{\sqrt{2^{d-|S|}}} \prod_{j\in S} \phi_{m_j}, \tag{25}$$

with the convention $\psi_\emptyset \coloneqq 1/\sqrt{2^d}$. Then the family

$$\left\{ \psi_S^{(\boldsymbol{m}_S)} \;:\; S \subseteq [d],\ \boldsymbol{m}_S \in \mathbb{N}_+^{|S|} \right\} \tag{26}$$

is a Hilbert basis of $L^2(\lambda)$. Indeed, for any $\boldsymbol{\alpha} \in \mathbb{N}^d$, let $S_{\boldsymbol{\alpha}} \coloneqq \{j \in [d] : \alpha_j > 0\}$ denote its support and $\boldsymbol{\alpha}_{S_{\boldsymbol{\alpha}}} \in \mathbb{N}_+^{|S_{\boldsymbol{\alpha}}|}$ its restriction to the nonzero coordinates. Since $\phi_0 = 1/\sqrt{2}$, one has

$$\Phi_{\boldsymbol{\alpha}} = \prod_{j=1}^{d} \phi_{\alpha_j} = \frac{1}{\sqrt{2^{d-|S_{\boldsymbol{\alpha}}|}}} \prod_{j\in S_{\boldsymbol{\alpha}}} \phi_{\alpha_j} = \psi_{S_{\boldsymbol{\alpha}}}^{(\boldsymbol{\alpha}_{S_{\boldsymbol{\alpha}}})}. \tag{27}$$

Remark A.5. A key consequence of the orthogonality of the basis is the orthogonality to the constant:

$$\forall S \subseteq [d] \setminus \emptyset,\ \forall \boldsymbol{m}_S \in \mathbb{N}_+^{|S|}, \quad \int_{[-1,1]^d} \psi_S^{(\boldsymbol{m}_S)}(\mathbf{x})d\lambda(\mathbf{x}) = 0. \tag{28}$$

B PROOFS

B.1 PROOF OF THEOREM 3.2

Lemma B.1. *The space $L^2([-1,1]^d, \lambda)$, that we will just denote $L^2(\lambda)$ for simplicity, is equivalent in norm to the space $L^2(P_{\mathbf{X}})$.*

Proof. Let $g : [-1,1]^d \to \mathbb{R}$ be a $P_{\mathbf{X}}-$square integrable function. By definition of $P_{\mathbf{X}}$ one has:

$$\|g\|^2_{L^2(P_{\mathbf{X}})} \;=\; \int_{[-1,1]^d} g(\mathbf{x})^2 p(\mathbf{x})\, d\lambda(\mathbf{x}). \tag{29}$$

Using that $c_1 \leq p \leq c_2$, we obtain:

$$c_1 \int_{[-1,1]^d} g(\mathbf{x})^2\, d\lambda(\mathbf{x}) \;\leq\; \int_{[-1,1]^d} g(\mathbf{x})^2 p(\mathbf{x})\, d\lambda(\mathbf{x}) \;\leq\; c_2 \int_{[-1,1]^d} g(\mathbf{x})^2\, d\lambda(\mathbf{x}),$$

which finally gives the norm equivalence. □

Definition B.2. For every set $S \subseteq [d]$, let Π_S be the orthogonal projection onto

$$V_S \coloneqq \overline{\text{span}} \left\{ \psi_S^{(\boldsymbol{m}_S)} \;\middle|\; \boldsymbol{m}_S \in \mathbb{N}_+^{|S|} \right\}, \tag{30}$$

where $\overline{\text{span}}$ denotes the closure of the linear span in $L^2(\lambda)$. Let also $\{w_S\}_{S\subseteq[d]}$ be a given collection of measurable functions, satisfying for all $S \subseteq [d]$ and all $\mathbf{x}_S \in [-1,1]^{|S|}$:

$$0 < \kappa_1 \leq w_S(\mathbf{x}_S) \leq \kappa_2 < \infty, \tag{31}$$

for a given couple of positive constants $0 < \kappa_1 \leq \kappa_2 < \infty$. We define the operator $\mathcal{T}$ as follows:

$$\mathcal{T} : \begin{cases} L^2(\lambda) & \longrightarrow & L^2(\lambda) \\ u & \longmapsto & \displaystyle\sum_{S\subseteq[d]} w_S\, \Pi_S(u), \end{cases} \tag{32}$$

Note that $\mathcal{T}$ depends on the given collection of functions $\{w_S\}_{S\subseteq[d]}$ but we simply denote it by $\mathcal{T}$ for simplicity.

Proposition B.3. *The operator $\mathcal{T}$ is bounded.*

Proof. Using that $\left(\psi_S^{(\boldsymbol{m}_S)}\right)_{S\subseteq[d],\, \boldsymbol{m}_S\in\mathbb{N}_+^{|S|}}$ is a Hilbert basis of $L^2(\lambda)$, we have the orthogonal decomposition

$$L^2(\lambda) = \bigoplus_{S\subseteq[d]}^{\perp} V_S. \tag{33}$$

Consequently, any $u \in L^2(\lambda)$ admits the unique orthogonal decomposition

$$u = \sum_{S\subseteq[d]} u_S, \qquad u_S \coloneqq \Pi_S(u). \tag{34}$$

Let $u \in L^2(\lambda)$, we have:

$$\|\mathcal{T}(u)\|_{L^2(\lambda)} = \left\| \sum_{S\subseteq[d]} w_S u_S \right\| \tag{35}$$

$$\leq \sum_{S\subseteq[d]} \|w_S u_S\| \tag{36}$$

$$\leq \kappa_2 \sum_{S\subseteq[d]} \|u_S\| \tag{37}$$

$$\leq \kappa_2 2^{d/2} \|u\| \tag{38}$$

which gives the desired result. □

Proposition B.4. *The operator $\mathcal{T}$ is invertible.*

Proof. For every $S, S' \subseteq [d]$, we define the following operator:

$$A_{S,S'} : \begin{cases} V_S & \longrightarrow & V_{S'} \\ u_S & \longmapsto & (\Pi_{S'} \circ \mathcal{T})(u_S) \end{cases} \tag{39}$$

We have the following identity for every $u \in L^2(\lambda)$:

$$\mathcal{T}(u) = \sum_{S \subseteq [d]} \sum_{S' \subseteq [d]} A_{S,S'}(\Pi_S(u)). \tag{40}$$

We denote by $L^2_S(\lambda)$ the Hilbert space of square-integrable functions depending only on the variables indexed by $S \subseteq [d]$. Observe that for every $S \subseteq [d]$, the operator $\mathcal{T}$ leaves this space invariant, i.e.,

$$\mathcal{T}\left(L^2_S(\lambda)\right) \subseteq L^2_S(\lambda). \tag{41}$$

Indeed, let $S \subseteq [d]$, the space $L^2_S(\lambda)$ admits the following orthogonal decomposition:

$$L^2_S(\lambda) = \overset{\perp}{\bigoplus_{T \subseteq S}} V_T. \tag{42}$$

So for any $u \in L^2_S(\lambda)$, u uniquely expresses as $u = \sum_{T \subseteq S} \Pi_T(u)$. By applying $\mathcal{T}$ to this sum, we obtain:

$$\mathcal{T}(u) = \sum_{T \subseteq S} w_T \Pi_T(u), \tag{43}$$

which is obviously an element of $L^2_S(\lambda)$. Furthermore, observe that we have the following decomposition:

$$L^2(\lambda) = \underbrace{\overset{\perp}{\bigoplus_{T \subseteq S}} V_T}_{L^2_S(\lambda)} \bigoplus \overset{\perp}{\bigoplus_{S' \not\subseteq S}} V_{S'} \tag{44}$$

A direct consequence of this stabilization property using the previous decomposition is:

$$\forall S, S' \subseteq [d], \quad S' \not\subseteq S \implies A_{S,S'} = 0, \tag{45}$$

Consequently, the operator matrix $\mathbf{A} := (A_{S,S'})_{S,S' \subseteq [d]}$ is triangular (an illustration in dimension 3 is provided in (55)). To show the invertibility of $\mathcal{T}$, it suffices to prove that all the diagonal blocs are invertible Tretter (2008), indeed in any finite partially-ordered family of blocks where off-diagonal blocks vanish outside $S' \subseteq S$, invertibility follows by induction on $|S|$. Let $S \subseteq [d]$, we will show that $A_{S,S}$ is invertible. First, $A_{S,S}$ is clearly a linear operator and so an endomorphism of V_S. Let take $u_S \in V_S$, we have:

$$\begin{aligned} \langle A_{S,S}(u_S), u_S \rangle &= \langle (\Pi_S \circ \mathcal{T})(u_S), u_S \rangle && (46) \\ &= \langle \Pi_S(w_S u_S), u_S \rangle && (47) \\ &= \langle w_S u_S, \Pi^\star_S(u_S) \rangle && (48) \\ &= \langle w_S u_S, \Pi_S(u_S) \rangle && (49) \\ &= \langle w_S u_S, u_S \rangle && (50) \\ &= \int w_S {u_S}^2 d\lambda && (51) \\ &\geq \kappa_1 \|u_S\|^2_{L^2(\lambda)}, && (52) \end{aligned}$$

so the operator $A_{S,S}$ is *coercive*. A similar computation would show that the operator is *bounded*, *i.e.*

$$\forall u_S, v_S \in V_S, \quad |\langle A_{S,S}(u_S), v_S \rangle| \leq \kappa_2 \|u_S\|_{L^2(\lambda)} \|v_S\|_{L^2(\lambda)}. \tag{53}$$

Finally, by applying the Lax-Milgram theorem (Brézis, 2011), we obtain that $A_{S,S}$ is invertible and the desired result. □

Remark B.5. Let $d = 3$. We illustrate the triangular structure of the matrix $\mathbf{A}$ and of the operator $\mathcal{T}$. Order the subsets of $[d]$ as follows:

$$\{\emptyset,\ \{1\},\ \{2\},\ \{3\},\ \{1,2\},\ \{1,3\},\ \{2,3\},\ \{1,2,3\}\}\,. \tag{54}$$

The coefficient $A_{S,S'}(u_S)$ denotes the projection onto $V_{S'}$ of $w_S u_S$, which belongs to $L^2_S(\lambda)$. In the case $d = 3$, the matrix takes the form

$$\begin{array}{c} \\ \emptyset \\ \{1\} \\ \{2\} \\ \{3\} \\ \{1,2\} \\ \{1,3\} \\ \{2,3\} \\ \{1,2,3\} \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & \{1\} & \{2\} & \{3\} & \{1,2\} & \{1,3\} & \{2,3\} & \{1,2,3\} \end{array} \\ \begin{bmatrix} * & & & & & & & \\ * & * & & & & & & \\ * & & * & & & & & \\ * & & & * & & & & \\ * & * & * & & * & & & \\ * & * & & * & & * & & \\ * & & * & * & & & * & \\ * & * & * & * & * & * & * & * \end{bmatrix} \end{array} \tag{55}$$

where $*$ indicates entries that are potentially non-zero.

Proposition B.6. *$\mathcal{T}$ is a linear homeomorphism of $L^2(\lambda)$.*

Proof. As a direct consequence of proposition B.3 and proposition B.4, $\mathcal{T}$ is a bounded automorphism of $L^2(\lambda)$. Using the Banach-Schauder theorem (Brézis, 2011), $\mathcal{T}$ is a homeomorphism. □

Proof of Theorem 3.2. First, observe that since $c_1 \leq p \leq c_2$, we have for all sets $S \subseteq [d]$:

$$c_1 2^{d-|S|} \leq p_S \leq c_2 2^{d-|S|}. \tag{56}$$

Consequently, a uniform bound for all p_S is given by:

$$c_1 \leq p_S \leq c_2 2^d. \tag{57}$$

Moreover, recall that our proposed collection of functions Ξ is given by:

$$\forall \mathbf{x} \in [-1,1]^d, \quad \Xi(\mathbf{x}) = \left(\left(\frac{\psi_S^{(\boldsymbol{m}_S)}(\mathbf{x})}{p_S(\mathbf{x}_S)} \right)_{\boldsymbol{m}_S \in \mathbb{N}_+^{|S|}} \right)_{S \subseteq [d]}. \tag{58}$$

So we can rewrite:

$$\forall S \subseteq [d], \forall \boldsymbol{m}_S \in \mathbb{N}_+^{|S|}, \quad \xi_S^{(\boldsymbol{m}_S)} = \mathcal{T}(\psi_S^{(\boldsymbol{m}_S)}), \tag{59}$$

by defining the weight functions w_S as follows: $w_S \coloneqq 1/p_S$ with $\kappa_1 \coloneqq 1/(c_2 2^d)$ and $\kappa_2 \coloneqq 1/c_1$. Thanks to Proposition B.6, Ξ is the direct image of a Hilbert basis by a homeomorphism $\mathcal{T}$ which implies that Ξ is a *Riesz basis* of $L^2(\lambda)$. In particular, by Lemma B.1 the family Ξ is also a Riesz basis of $L^2(P_{\mathbf{X}})$ which gives the desired conclusion. □

B.2 Proof of Lemma 3.3

Proof. Let us consider $T \subsetneq S \subseteq [d]$, $\boldsymbol{m}_S := (m_s)_{s\in S} \in \mathbb{N}_+^{|S|}$, $\boldsymbol{n}_T := (n_t)_{t\in T} \in \mathbb{N}_+^{|T|}$. For simplicity, we define $C := S \setminus T$ and consider $\boldsymbol{m}_C := (m_c)_{c\in C}$. We have:

$$\mathbb{E}\left[\xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}) \cdot \xi_T^{(\boldsymbol{n}_T)}(\mathbf{X})\right] = \int_{[-1,1]^d} \xi_S^{(\boldsymbol{m}_S)}(\mathbf{x}) \cdot \xi_T^{(\boldsymbol{n}_T)}(\mathbf{x}) \cdot p(\mathbf{x}) d\lambda(\mathbf{x}), \tag{60}$$

$$= \int_{[-1,1]^{|S|}} \xi_S^{(\boldsymbol{m}_S)}(\mathbf{x}) \cdot \xi_T^{(\boldsymbol{n}_T)}(\mathbf{x}) \cdot p_S(\mathbf{x}_S) d\lambda(\mathbf{x}_S), \tag{61}$$

$$\propto \int_{[-1,1]^{|S|}} \frac{\prod_{s\in S} \phi_{m_s}(\mathbf{x}_s)}{p_S(\mathbf{x}_S)} \frac{\prod_{t\in T} \phi_{n_t}(\mathbf{x}_t)}{p_T(\mathbf{x}_T)} p_S(\mathbf{x}_S) d\lambda(\mathbf{x}_S), \tag{62}$$

$$\propto \int_{[-1,1]^{|S|}} \prod_{s\in S} \phi_{m_s}(\mathbf{x}_s) \frac{\prod_{t\in T} \phi_{n_t}(\mathbf{x}_t)}{p_T(\mathbf{x}_T)} d\lambda(\mathbf{x}_S), \tag{63}$$

$$\propto \int_{[-1,1]^{|S|}} \underbrace{\prod_{c\in C} \phi_{m_c}(\mathbf{x}_c)}_{\sqrt{2^{d-|C|}}\psi_C^{(\boldsymbol{m}_C)}(\mathbf{x}_C)} \cdot \underbrace{\frac{\prod_{t\in T} \phi_{m_t}(\mathbf{x}_t)\phi_{n_t}(\mathbf{x}_t)}{p_T(\mathbf{x}_T)}}_{u(\mathbf{x}_T)} d\lambda(\mathbf{x}_S), \tag{64}$$

$$\propto \underbrace{\int_{[-1,1]^{|C|}} \psi_C^{(\boldsymbol{m}_C)}(\mathbf{x}_C) d\lambda(\mathbf{x}_C)}_{=0} \cdot \int_{[-1,1]^{|T|}} u(\mathbf{x}_T) d\lambda(\mathbf{x}_T), \tag{65}$$

$$= 0. \tag{66}$$

□

Remark B.7. In particular, all basis elements are centered, *i.e.*

$$\forall S \subseteq [d] \setminus \emptyset, \quad \mathbb{E}\left[\xi_S^{(\boldsymbol{m}_S)}(\mathbf{X})\right] = 0. \tag{67}$$

B.3 Proof of Theorem 3.4

Proof of Theorem 3.4. By Theorem 3.2, every function $f \in L^2(P_{\mathbf{X}})$ admits a unique representation

$$f(\mathbf{X}) = \sum_{S\subseteq[d]} \sum_{\boldsymbol{m}_S \in \mathbb{N}_+^{|S|}} \theta_S^{(\boldsymbol{m}_S)} \xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}), \tag{68}$$

where the coefficient sequence

$$\boldsymbol{\theta} := \left(\theta_S^{(\boldsymbol{m}_S)}\right)_{S\subseteq[d],\, \boldsymbol{m}_S \in \mathbb{N}_+^{|S|}}$$

is uniquely determined by f and belongs to ℓ^2.

For every $S \subseteq [d]$, define

$$\mathcal{H}_S := \overline{\operatorname{span}}\left\{ \xi_S^{(\boldsymbol{m}_S)} \;\middle|\; \boldsymbol{m}_S \in \mathbb{N}_+^{|S|} \right\}. \tag{69}$$

Since Ξ is a Riesz basis, its expansion converges unconditionally. Consequently, the series can be regrouped according to the subsets S, and

$$L^2(P_{\mathbf{X}}) = \bigoplus_{S\subseteq[d]} \mathcal{H}_S, \tag{70}$$

where the right-hand side denotes an algebraic direct sum, which is not orthogonal in general.

Moreover, if $T \subsetneq S$, Lemma 3.3 shows that every basis element of $\mathcal{H}_S$ is orthogonal to every basis element of $\mathcal{H}_T$. By linearity and continuity of the inner product, this extends to the closures, so that

$$T \subsetneq S \quad \Longrightarrow \quad \mathcal{H}_T \perp \mathcal{H}_S. \tag{71}$$

For every $T \subseteq [d]$, the same Riesz-basis decomposition applied to functions of $\mathbf{X}_T$ yields

$$L_T^2 = \bigoplus_{U \subseteq T} \mathcal{H}_U. \tag{72}$$

Hence, if $T \subsetneq S$, then every $U \subseteq T$ is a strict subset of S, and therefore

$$\mathcal{H}_S \perp L_T^2. \tag{73}$$

Finally, define

$$f_S := \sum_{\boldsymbol{m}_S \in \mathbb{N}_+^{|S|}} \theta_S^{(\boldsymbol{m}_S)} \xi_S^{(\boldsymbol{m}_S)}. \tag{74}$$

By definition, we have $f = \sum_{S \subseteq [d]} f_S$, while the previous argument shows that

$$\langle f_S, g_T \rangle_{L^2(P_{\mathbf{X}})} = 0 \qquad \text{for every } T \subsetneq S \text{ and every } g_T \in L_T^2.$$

Thus the family $(f_S)_{S \subseteq [d]}$ satisfies the ANOVA decomposition (2) and the hierarchical orthogonality constraint (3) which completes the proof. □

B.4 Proof of Corollary 3.5

Proof. To show that the resulting decomposition is mutually orthogonal by set, it suffices to prove that for any sets S, T such that $S \neq T$ and for any vector of indices $\boldsymbol{m}_S, \boldsymbol{n}_T$, one has:

$$\mathbb{E}\left[\xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}) \cdot \xi_T^{(\boldsymbol{n}_T)}(\mathbf{X})\right] = 0. \tag{75}$$

Under independence, the density expresses as the following product $p = p^{\otimes d} := p_1 \times \cdots \times p_d$. The elements of our basis becomes:

$$\forall S \subseteq [d], \forall \boldsymbol{m}_S \in \mathbb{N}_+^{|S|}, \quad \xi_S^{(\boldsymbol{m}_S)} = \frac{1}{\sqrt{2^{d-|S|}}} \cdot \prod_{j \in S} \frac{\phi_{m_j}}{p_j}. \tag{76}$$

Let $S \neq T \subseteq [d]$, without loss of generality, we fix $s^\star \in S$ such that $s^\star \notin T$. We have:

$$\mathbb{E}\left[\xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}) \cdot \xi_T^{(\boldsymbol{n}_T)}(\mathbf{X})\right] = \mathbb{E}\left[\frac{1}{\sqrt{2^{d-|S|}}} \prod_{s \in S} \frac{\phi_{m_s}(X_s)}{p_s(X_s)} \cdot \frac{1}{\sqrt{2^{d-|T|}}} \prod_{t \in T} \frac{\phi_{n_t}(X_t)}{p_t(X_t)}\right] \tag{77}$$

$$\propto \mathbb{E}\left[\prod_{s \in S} \frac{\phi_{m_s}(X_s)}{p_s(X_s)} \cdot \prod_{t \in T} \frac{\phi_{n_t}(X_t)}{p_t(X_t)}\right] \tag{78}$$

$$\propto \mathbb{E}\left[\underbrace{\frac{\phi_{m_{s^\star}}(X_{s^\star})}{p_{s^\star}(X_{s^\star})}}_{u(X_{s^\star})} \cdot \underbrace{\prod_{s \in S \setminus \{s^\star\}} \frac{\phi_{m_s}(X_s)}{p_s(X_s)} \cdot \prod_{t \in T} \frac{\phi_{n_t}(X_t)}{p_t(X_t)}}_{v(X_{[d]\setminus\{s^\star\}})}\right] \tag{79}$$

$$\propto \mathbb{E}\left[\frac{\phi_{m_{s^\star}}(X_{s^\star})}{p_{s^\star}(X_{s^\star})}\right] \cdot \mathbb{E}\left[\prod_{s \in S \setminus \{s^\star\}} \frac{\phi_{m_s}(X_s)}{p_s(X_s)} \cdot \prod_{t \in T} \frac{\phi_{n_t}(X_t)}{p_t(X_t)}\right] \tag{80}$$

$$\propto \mathbb{E}\left[\frac{\phi_{m_{s^\star}}(X_{s^\star})}{p_{s^\star}(X_{s^\star})}\right] \tag{81}$$

$$\propto \int_{-1}^{1} \frac{\phi_{m_{s^\star}}(x)}{p_{s^\star}(x)} p_{s^\star}(x) dx \tag{82}$$

$$\propto \int_{-1}^{1} \phi_{m_{s^\star}}(x) dx \tag{83}$$

$$= 0. \tag{84}$$

□

C ESTIMATION

In this section, we detail all the computations behind our method to estimate the functional ANOVA from a mere data sample. Throughout this section, we denote by $\mathcal{X} := \{\mathbf{X}^{(i)}\}_{i=1}^n$ and $\mathcal{Y} := \{f(\mathbf{X}^{(i)})\}_{i=1}^n$ the data, and $\{\mathcal{X}, \mathcal{Y}\}$ is the corresponding data sample. If in the vast majority of datasets, one can assume that the features are bounded, they rarely live natively in $[-1, 1]^d$. We illustrate in Figure 3 the entire procedure to estimating the truncated functional decomposition from data.

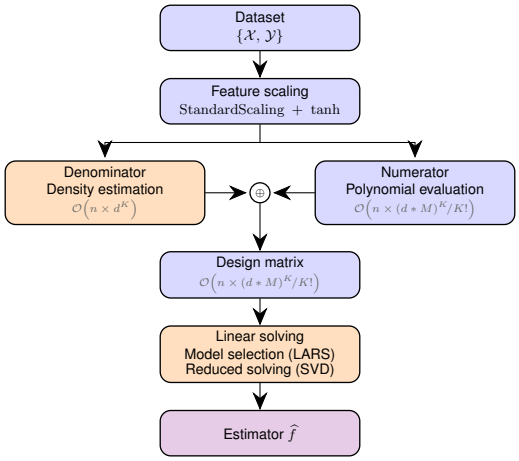


Figure 3: Illustration of our method to estimate the decomposition *from scratch* (see Appendix C).

Feature scaling. The first step to apply our framework on a real world dataset is a feature scaling in $[-1, 1]$. We propose to address this by first centering and reducing our data and then applying $\tanh$ to each feature. This can be described as follows:

$$\mathcal{X} \xrightarrow{\text{StandardScaling}} \mathcal{X}_{(0,1)} \xrightarrow{\tanh} \mathcal{X}_{\text{scaled}} \tag{85}$$

Features are standardized to zero mean and unit variance, using the `StandardScaler` from `scikit-learn` Pedregosa et al. (2011).

Density estimation. We follow the standard projection procedure to estimate a density function using an orthonormal basis, see for example Cencov (1962) and Tsybakov (2009). From now on, we assume that the feature sample takes values in $[-1, 1]^d$. Our theoretical guarantees depend on the true marginal densities p_S for $S \subseteq [d]$, which are unavailable in practice and must be estimated from the data. We propose to also estimate each p_S using a truncated tensorized Legendre expansion. In practice, for any $S \subseteq [d]$ and a truncation degree d_{density}, we approximate p_S as follows:

$$p_S \approx \sum_{\boldsymbol{m} \in \{0, \dots, d_{\text{density}}\}^{|S|}} c_S^{(\boldsymbol{m})} \prod_{j \in S} \phi_{m_j}, \tag{86}$$

By orthonormality of the tensorized basis, each coefficient $c_S^{(\boldsymbol{m})}$ is estimated as follows:

$$\widehat{c}_{S,n}^{(\boldsymbol{m})} := \frac{1}{n} \sum_{i=1}^n \prod_{j \in S} \phi_{m_j}(X_j^{(i)}). \tag{87}$$

This choice is motivated by two considerations. First, it keeps the whole pipeline fully transparent and free of additional black-box components: each coefficient reduces to an empirical average, the estimator is trained from scratch with a single hyperparameter (the truncation degree d_{density}), and

it integrates seamlessly with the rest of our analysis, which is already expressed in the Legendre basis. Second, and more importantly, our goal is not to recover p_S pointwise but to enforce the hierarchical orthogonality constraint (3); since capturing a *global trend* of p_S is enough to guarantee this condition, the pointwise quality of the density estimate is largely irrelevant to our objective.

Remark C.1. In theory the number of densities to be estimated scales as $\mathcal{O}(d^K)$, but in practice we restrict ourselves to $K \in \{1, 2\}$, which keeps the estimation step tractable even in moderately high dimension.

Remark C.2 (Density clipping). In practice, we replace the raw estimate $\hat{p}_S$ by $\max(\varepsilon, \hat{p}_S)$ (typically $\varepsilon = 10^{-2}$), in order to avoid numerical instabilities caused by regions where the estimated density is very close to zero (or negative, since the projection estimator is not constrained to be nonnegative).

Legendre polynomials are evaluated using `eval_legendre` from `SciPy` Virtanen et al. (2020). The truncation degree d_{density} and the clipping parameter ε constitute two hyperparameters of the method, referred to as `deg_density` and `density_clip` respectively.

Design matrix. With a slight abuse of notation, we still write p_S for the clipped estimated densities. We then form the *design matrix* associated with our basis functions,

$$\mathbf{B} \;:=\; \Big(\xi_S^{(\boldsymbol{m}_S)}(\mathbf{X}^{(i)})\Big)_{\substack{i\in\{1,\dots,n\}\\(S,\boldsymbol{m}_S)\in\mathcal{I}_K^M}} \;\in\; \mathbb{R}^{n\times\mathcal{N}(d,K,M)}. \tag{88}$$

The numerator of the design matrix is similarly computed via Legendre polynomials using `eval_legendre` from `SciPy`. The resulting matrix $\mathbf{B}$ is stored as a two-dimensional `NumPy` array.

Linear solving. The estimator defined in (18) can now be expressed as follows:

$$\widehat{\boldsymbol{\theta}}_n^{K,M} \;\in\; \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{\mathcal{N}(d,K,M)}} \|\mathbf{B}\,\boldsymbol{\beta} \;-\; \boldsymbol{y}\|_2^2, \qquad \boldsymbol{y} \;:=\; \Big(f(\mathbf{X}^{(i)})\Big)_{1\le i\le n} \in \mathbb{R}^n, \tag{89}$$

where $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^n$. We proceed in two stages: we first select a relevant subset of columns via a LARS procedure Efron et al. (2004); Blatman and Sudret (2011), and then solve the resulting reduced system using an SVD.

Model selection is carried out using `LassoLarsIC` from `scikit-learn`, which traverses the LARS regularization path and selects the optimal support via the Bayesian Information Criterion (BIC). This yields a subset of column indices defining the reduced design matrix $\mathbf{B}_{\text{red}}$. The corresponding least-squares system is then solved via `numpy.linalg.lstsq` Harris et al. (2020), which computes the minimum-norm solution through a thin SVD decomposition.

Estimator. We denote by $\widehat{f}$ the resulting estimator, with $\widehat{f}_S$ its component associated to each subset $S \subseteq [d]$. The full decomposition over the dataset denoted `X, y` is obtained by calling a function of the following arguments:

```
K, M, d_density, density_clip, X, y.
```

Remark C.3. As a final step, we recenter the resulting ANOVA components so that $f_\emptyset$ captures the global mean. This operation comes at negligible computational cost and merely translates the components, leaving the structure of Ξ and the hierarchical orthogonality unchanged. Prior to this step, the components were already nearly centered by construction; the residual offset is an empirical bias due to finite-sample approximation error, and correcting for it at essentially no cost is therefore worthwhile.

Remark C.4. The $\tanh$ transformation, applied component-wise, maps the features into $[-1, 1]^d$, as required by our framework. We emphasize that this transformation is applied *solely within our explainer*: the predictive model is trained on the original data, and the baseline explainers we compare against likewise operate on the untransformed data. The comparison is therefore fair, since the

underlying function of interest is identical across all methods; only its decomposition is computed in a different coordinate system.

Moreover, by uniqueness of the functional ANOVA decomposition, this change of variables leaves the importance attributions invariant. Let $\mathbf{Z}$ denote the original features and $\mathbf{X} \coloneqq \tanh(\mathbf{Z})$ the corresponding transformation with $\tanh$ applied component-wise. Since $\tanh$ is a bijection, the marginal structure is preserved. Let f be the function of interest defined on the original space, with the unique decomposition

$$f(\mathbf{Z}) = \sum_{S \subseteq [d]} f_S(\mathbf{Z}_S), \tag{90}$$

and define

$$\widetilde{f}(\mathbf{X}) \coloneqq f\left(\tanh^{-1}(\mathbf{X})\right), \tag{91}$$

which in turn admits a unique decomposition $\widetilde{f}(\mathbf{X}) = \sum_{S \subseteq [d]} \widetilde{f}_S(\mathbf{X}_S)$. Substituting $\mathbf{X} = \tanh(\mathbf{Z})$ yields

$$f(\mathbf{Z}) \;=\; \widetilde{f}(\tanh(\mathbf{Z})) \;=\; \sum_{S \subseteq [d]} \widetilde{f}_S(\tanh(\mathbf{Z}_S))\,. \tag{92}$$

By uniqueness of this decomposition, we conclude that:

$$\forall S \subseteq [d], \quad \widetilde{f}_S(\mathbf{X}_S) = f_S(\mathbf{Z}_S). \tag{93}$$

D Experiments details

D.1 Resources

All experiments were conducted on a MacBook Pro M4 with 32 GB of RAM. The code is written in `Python` using mainly `numpy`, `scipy`, and `scikit-learn`. All computations reduce to standard linear-algebra operations, primarily array manipulation and the resolution of linear systems and are executed on CPU.

D.2 Machine learning models

We describe the four machine learning models used in the numerical experiments of this paper. The same architectures are used consistently across all experiments, including both the analytical setting and the real-world datasets. All of these machine learning models are treated as *black boxes* by our method.

XGBoost. We use `XGBoost` Chen et al. (2015), a gradient-boosted decision tree ensemble, with 100 estimators, a maximum depth of 10, and a learning rate of 0.05. Both row and column subsampling are set to 0.8 at each boosting iteration to mitigate overfitting. We use an $80/20$ train-test split, with early stopping after 30 rounds without improvement on the held-out set. For classification tasks, the split is stratified and the objective is set to `binary:logistic` or `multi:softprob` depending on the number of classes; for regression, the default squared-error objective is used. The random seed is exposed as a pipeline parameter.

MLP. We use a standard feed-forward multilayer perceptron with three hidden layers of widths $(128, 64, 32)$ and ReLU activations, followed by a dropout rate of 0.1 applied after each hidden layer. Features are standardized using a `StandardScaler` fitted on the training set. The model is trained with the Adam optimizer Kingma and Ba (2014), a learning rate of 10^{-3}, a weight decay of 10^{-4}, a batch size of 256, and a maximum of 500 epochs, with early stopping after 30 epochs without improvement on the held-out validation set. The loss function is set to mean squared error for regression, binary cross-entropy with logits for two-class classification, and categorical cross-entropy otherwise. We use an $80/20$ train-test split, stratified in the classification setting. The random seed is exposed as a pipeline parameter.

NAM. Neural Additive Models Agarwal et al. (2021) fit a generalized additive model of the form $h(\mathbf{x}) = C + \sum_{j=1}^{d} h_j(\mathbf{x}_j)$, where each *shape function* h_j is parameterized by an independent univariate subnetwork. Following the standard ReLU variant of Agarwal et al. (2021), each subnetwork consists of three hidden layers of widths $(64, 64, 32)$ with ReLU activations. Features are standardized on the training set prior to training. The model is trained with the Adam optimizer Kingma and Ba (2014), a learning rate of 6.74×10^{-3} decayed exponentially by a factor of 0.995 per epoch, weight decay of 10^{-6}, and an additional L_2 penalty of 10^{-3} on the shape function outputs to encourage smoothness. We use a batch size of 1024, a maximum of 1000 epochs, and early stopping after 50 epochs without improvement on the held-out validation set. The loss function is mean squared error for regression, binary cross-entropy with logits for two-class classification, and categorical cross-entropy otherwise. After training, the shape functions are centered so that each h_j has zero empirical mean over the training set, with the residual absorbed in the intercept C. We use an $80/20$ train-test split, stratified in the classification setting, and the random seed is exposed as a pipeline parameter.

EBM. Explainable Boosting Machines Nori et al. (2019) are second-order generalized additive models of the form $h(\mathbf{x}) = C + \sum_j h_j(\mathbf{x}_j) + \sum_{j,k} h_{j,k}(\mathbf{x}_j, \mathbf{x}_k)$, where each univariate and bivariate function is fitted greedily using cyclic gradient boosting with shallow decision trees. We use the reference implementation provided by the `interpret` package Nori et al. (2019), with a maximum of 256 bins for feature discretization, 10 automatically selected pairwise interactions, and up to $5\,000$ boosting rounds with a learning rate of 10^{-2}. Each base learner is a shallow tree with at most 3 leaves and a minimum of 2 samples per leaf. We use an $80/20$ train-test split, stratified in the classification setting. The random seed is exposed as a pipeline parameter.

D.3 Explainers

We fix in advance an evaluation set $\mathcal{X}_{\text{eval}}$ of shape $N_{\text{eval}} \times d$ on which all subsequent quantities are computed and compared. In the analytical setting, we explain the entire dataset, i.e., $N_{\text{eval}} = n = 10\,000$. For the real-world datasets, we sample $N_{\text{eval}} = 5\,000$ instances uniformly at random. For a given dataset, all methods are evaluated on the same set of instances to ensure a fair comparison.

KernelSHAP. We fix in advance a background dataset of size N_{bg} and apply the `KernelExplainer` from the `shap`[1] package Lundberg and Lee (2017) directly to the trained model. In the analytical setting, we apply this method directly to the function f, treated as a black box. For the real-world datasets, it is applied exclusively to the trained MLP.

DeepSHAP. We fix in advance a background dataset of size N_{bg} and apply the `DeepExplainer` from the `shap` package Lundberg and Lee (2017) directly to the trained model. For the real-world datasets, it is applied exclusively to the trained MLP.

Remark D.1. In practice, we set $N_{\text{bg}} = 200$ for the datasets *Bike Sharing*, *California Housing* and *Pima Indians Diabetes*, a standard value in the literature. For the datasets *Adult Census Income*, *Electrical Grid Stability* and *Superconductivity*, we use $N_{\text{bg}} = 50$ for computational reasons.

TreeSHAP. We apply TreeSHAP Lundberg et al. (2018), via the `TreeExplainer` class of the `shap` package, to the trained XGBoost model. We use the `tree_path_dependent` option, which computes Shapley values in the observational sense by accounting for feature dependencies through the tree structure.

TreeHFD. We apply TreeHFD Bénard (2025), via the `treehfd`[2] package, to the same trained XGBoost model as TreeSHAP. We use the first-order mode (`interaction_order=1`), which restricts the decomposition to main effects, both for computational efficiency and for consistency with our figures, in which we display the main effects.

EBM. The additive decomposition is part of the fitted EBM itself. For each evaluation instance $\mathbf{x}$ and each feature j, the main effect $h_j(\mathbf{x}_j)$ is extracted directly via the `eval_terms` method of the

[1] https://github.com/shap/shap

[2] https://github.com/ThalesGroup/treehfd

`interpretML`[3] package Nori et al. (2019), which returns the contribution of every univariate (and bivariate) shape function to the model output. Displayed main effects are obtained by plotting $h_j(x_j)$ against x_j over the evaluation set. In the classification setting, all contributions are reported on the logit scale.

NAM. Likewise, the additive decomposition is part of the NAM architecture. For each evaluation instance $\mathbf{x}$ and each feature j, the main effect $h_j(\mathbf{x}_j)$ is obtained by a forward pass of the j-th subnetwork on the scalar input x_j. After training, the shape functions are centered so that $\mathbb{E}[h_j(\mathbf{X}_j)] = 0$ on the training set, with the residual absorbed in the intercept C Agarwal et al. (2021). Displayed main effects are obtained as for EBM, by plotting $h_j(\mathbf{x}_j)$ against $\mathbf{x}_j$ over the evaluation set; in classification, they are reported on the logit scale.

D.4 Analytical Case

Theoretical decomposition. In this case, we consider a random vector $\mathbf{X} \in [-1,1]^3$ whose joint density belongs to the Farlie–Gumbel–Morgenstern (Nelsen, 2006) family and is given for $\mathbf{x} \in [-1,1]^3$ by:

$$p_\rho(\mathbf{x}) = \frac{1}{8}\big(1 + \rho\,(x_1x_2 + x_2x_3 + x_1x_3)\big), \tag{94}$$

where $\rho \in (-\frac{1}{3},\, 1)$ is the dependence parameter. The marginal densities of order 1 and 2 are explicit and are given by

$$p_j(x_j) = \frac{1}{2}, \tag{95}$$

$$p_{i,j}(x_i, x_j) = \frac{1}{4}\big(1 + \rho\, x_i\, x_j\big). \tag{96}$$

Each marginal is uniform on $[-1,1]$, and the pairwise correlation satisfies $\mathrm{Corr}(X_i, X_j) = \rho/3$ for all $i \neq j$. Note that for any $\rho \in (-1/3, 1)$, we have:

$$0 < \underbrace{\frac{1}{8}(1-\rho)}_{c_1} \leq p_\rho \leq \underbrace{\frac{1}{8}(1+3\rho)}_{c_2} < \infty, \tag{97}$$

which satisfies the Assumption 2.1. We define the theoretical components of f in terms of elements of Ξ as follows:

$$f_1(x_1) \coloneqq \frac{\sqrt{\frac{2}{7}}\,\phi_3(x_1) - \sqrt{\frac{2}{3}}\,\phi_1(x_1)}{p_1(x_1)}, \tag{98}$$

$$f_2(x_2) \coloneqq \frac{\sqrt{\frac{2}{3}}\,\phi_1(x_2) + \sqrt{\frac{2}{5}}\,\phi_2(x_2)}{p_2(x_2)}, \tag{99}$$

$$f_{1,2}(x_1, x_2) \coloneqq \frac{\frac{2}{9}\,\phi_4(x_1)\phi_4(x_2) + \frac{2}{17}\,\phi_8(x_1)\phi_8(x_2)}{p_{1,2}(x_1, x_2)}, \tag{100}$$

and $f_\emptyset = f_3 = f_{1,3} = f_{2,3} = f_{1,2,3} = 0$.

Simulations. In our theoretical setup, the function f is available in closed form, and samples of $\mathbf{X}$ can be drawn via rejection sampling from the uniform distribution on $[-1,1]^3$ using the density p_ρ. We set $\rho = 1/2$ for simplicity. The illustration in Fig 1 is produced with $n = 10\,000$ samples.

Comparisons. We compare the main effects f_1, f_2, f_3 against several explanation methods (KernelSHAP, TreeSHAP, TreeHFD) and interpretable models (NAM and EBM). In each case, we follow the procedure described above and explain the $10\,000$ instances of the dataset.

- **KernelSHAP.** We apply KernelSHAP directly to f, treated as a *black-box* model.
- **TreeSHAP.** We train an XGBoost model and apply TreeSHAP to the trained estimator.

[3] https://interpret.ml/

- **TreeHFD.** We apply TreeHFD to the same XGBoost model.
- **NAM.** We train a NAM and report its learned main effects.
- **EBM.** We train an EBM and report its learned main effects.

Remark D.2. Note that we apply our own estimator on this synthetic dataset, assuming that the theoretical f is black-box. And we verify that our method also captures the true underlying ANOVA structure (see Fig.4).

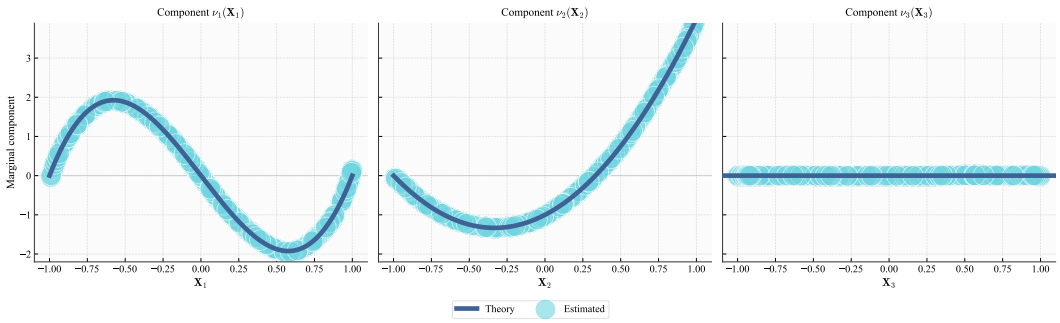


Figure 4: Estimated main effects in the analytical setting for X_1, X_2, and the irrelevant variable X_3. We take $K = 2$, $M = 10$, $d_{\text{density}} = 10$ and $\varepsilon = 0.01$

D.5 Datasets

The *California Housing* dataset is directly available from `scikit-learn`. The remaining datasets are retrieved from the UCI Machine Learning Repository Dua et al. (2017) using the `fetch_ucirepo` function of the `ucimlrepo`[4] package. All features are used for every dataset, with the exception of *Bike Sharing*, where we retain only eight: `hr`, `weekday`, `holiday`, `season`, `atemp`, `hum`, `windspeed`, and `weathersit` following Bénard (2025).

D.6 Hyperparameters

In this section, we specify the hyperparameters used by our method on each dataset. Recall that these hyperparameters are:

1. K: the truncation order of the interaction levels;
2. M: the maximum polynomial degree (numerator);
3. d_{density}: the maximum polynomial degree used to estimate the marginal densities (denominator);
4. ε: the clipping threshold applied to the estimated densities.

Table 3: Hyperparameters of our method for all models across all datasets. The – entries correspond to datasets on which the NAM baseline was not trained for computational reasons.

	XGBoost				MLP				EBM				NAM			
ID	K	M	d_{density}	ε	K	M	d_{density}	ε	K	M	d_{density}	ε	K	M	d_{density}	ε
BS	2	10	5	0.01	2	8	6	0.01	2	10	4	0.01	–	–	–	–
CH	2	10	4	0.01	1	10	4	0.01	2	10	4	0.01	1	10	4	0.01
CI	1	10	4	0.01	1	10	4	0.01	1	10	6	0.01	–	–	–	–
EG	2	4	4	0.01	2	4	4	0.01	2	4	4	0.01	1	4	4	0.01
PI	2	5	4	0.1	2	5	4	0.01	2	5	4	0.01	1	10	4	0.01
SC	1	10	10	0.01	1	10	10	0.01	1	10	10	0.01	–	–	–	–

[4] https://pypi.org/project/ucimlrepo/

Remark D.3. In practice, we always take $K \in \{1, 2\}$, so that our method reduces to a GAM ($K = 1$) or a GA^{2}M ($K = 2$). The density-estimation degree d_{density} can also be kept small: a low-degree polynomial is typically sufficient to capture the overall *trend* of the marginal densities. Finally, we never go beyond $M = 10$ for the shape-function polynomials, as we empirically observe no additional signal capture beyond this degree. Pushing M further can even degrade the hierarchical orthogonality of the decomposition, a signal that either a more sophisticated penalized model selection scheme or an increase of the truncation order to $K = 3$ would be required.

D.7 Performance metrics

In this subsection, we report all performance metrics that were omitted from the main body of the paper.

Tables. For clarity, we reproduce here the tables reporting the dataset characteristics and the performance of our method across the four models and all datasets. Tab 4 and Tab 5 report the predictive performance of the trained models on the six real-world datasets, together with the corresponding metrics of our method.

Table 4: Dataset characteristics and performance of our method.

			XGB				MLP			
ID	n	d	Acc.	K	Rec.	Ortho.	Acc.	K	Rec.	Ortho.
BS	17 379	8	0.86	2	0.92	9.49e−2	0.85	2	0.91	8.83e−2
CH	20 640	8	0.84	2	0.86	6.35e−2	0.81	1	0.88	0.00
CI	48 842	14	0.88	1	0.91	0.00	0.85	1	0.94	0.00
EG	10 000	12	0.93	2	0.89	7.41e−3	0.98	2	0.95	6.61e−3
PI	768	8	0.74	2	0.85	9.56e−2	0.73	2	0.99	0.00
SC	21 263	81	0.93	1	0.86	0.00	0.91	1	0.89	0.00

Table 5: Dataset characteristics and performance of our method on EBM and NAM.

			EBM				NAM			
ID	n	d	Acc.	K	Rec.	Ortho.	Acc.	K	Rec.	Ortho.
BS	17 379	8	0.83	2	0.96	9.52e−2	–	–	–	–
CH	20 640	8	0.83	2	0.89	4.04e−2	0.76	1	0.94	0.00
CI	48 842	14	0.88	1	0.90	0.00	–	–	–	–
EG	10 000	12	0.94	2	0.99	5.58e−3	0.88	1	1.00	0.00
PI	768	8	0.76	2	0.97	0.00	0.73	1	1.00	0.00
SC	21 263	81	0.90	1	0.88	0.00	–	–	–	–

Acc. denotes the predictive performance of the trained model: classification accuracy for classification tasks and the coefficient of determination for regression tasks. The quantity *Rec.* is the reconstruction coefficient of determination of our estimator, i.e. the fraction of the output variance captured by our decomposition on the entire dataset. Finally, *Ortho.* is defined as the following absolute cosine

$$\text{Ortho.} := \max_{S,T} \left| \cos\left(\widehat{f}_S, \widehat{f}_T\right) \right|. \tag{101}$$

The cosine is formally defined as follows:

$$\cos(U, V) := \frac{\langle U, V \rangle}{\|U\| \|V\|} = \frac{\mathbb{E}[UV]}{\sqrt{\mathbb{E}[U^2]}\sqrt{\mathbb{E}[V^2]}}. \tag{102}$$

The absolute cosine measures the deviation from orthogonality : the more the quantity is approaching zero, the more the *angle* is near to the orthogonality. Finally, for significance reasons, the max is taken over the following ensemble:

$$\{(S, T) : T \subsetneq S,\ \mathbb{V}(\widehat{f}_S(\mathbf{X}_S))/\mathbb{V}(f(\mathbf{X})) \geq 1\%\}. \tag{103}$$

We restrict the maximum to interaction components whose variance represents at least 1% of the output variance (following Bénard (2025)), so as to discard components with negligible contribution to the decomposition.

Computation time. The computation times of our method on each real-world dataset are reported in Tab. 6. For each dataset, we fix the random seed to 42 and train the corresponding machine learning model. We then apply our decomposition method to each trained model over 10 independent runs, and report the mean and standard deviation. We observe that our method is extremely fast, enabling the explanation of hundreds or thousands of instances in a few seconds or less. For instance, the quasi-additive structure of an MLP trained on the *Adult Census Income* dataset allows us to explain all 48 842 instances in approximately 0.6 s. We display in the Table 6 the computation time to obtain the entire truncated decomposition for each dataset.

Table 6: Time (in s) to estimate the decomposition on the entire dataset (Mean $\pm$ std over 10 repetitions).

ID	n	d	XGB	MLP	EBM	NAM
BS	17 379	8	9.806 ± 0.228	5.741 ± 0.182	9.721 ± 0.206	–
CH	20 640	8	11.048 ± 0.415	0.129 ± 0.012	11.077 ± 0.414	0.126 ± 0.005
CI	48 842	14	0.575 ± 0.023	0.597 ± 0.053	0.691 ± 0.010	–
EG	10 000	12	1.420 ± 0.048	1.407 ± 0.025	1.365 ± 0.019	0.039 ± 0.007
PI	768	8	0.237 ± 0.006	0.230 ± 0.009	0.230 ± 0.006	0.013 ± 0.001
SC	21 263	81	3.520 ± 0.097	3.547 ± 0.159	3.501 ± 0.053	–

Hierarchical orthogonality. As a consequence of the hierarchical orthogonality constraint (3), the components $\{f_S(\mathbf{X}_S)\}_{S\subseteq[d]}$ must satisfy

$$\forall S, T \subseteq [d], \quad T \subsetneq S \implies \mathbb{E}[f_S(\mathbf{X}_S)\, f_T(\mathbf{X}_T)] = 0. \tag{104}$$

In practice, we measure the empirical cosine between pairs of components f_S and f_T with $T \subsetneq S \subseteq [d]$. When $T = \emptyset$, the cosine is defined by:

$$\frac{\mathbb{E}[f_S(\mathbf{X}_S)]}{\sqrt{\mathbb{E}[f_S(\mathbf{X}_S)^2]}}, \tag{105}$$

which quantifies the degree of centering of f_S. For $T \neq \emptyset$, we compute

$$\frac{\mathbb{E}[f_S(\mathbf{X}_S)\, f_T(\mathbf{X}_T)]}{\sqrt{\mathbb{E}[f_S(\mathbf{X}_S)^2]\ \mathbb{E}[f_T(\mathbf{X}_T)^2]}}. \tag{106}$$

After estimation, we systematically recenter the estimated components $\widehat{f}_S$, so that for every $S \subseteq [d]$,

$$\frac{\mathbb{E}\left[\widehat{f}_S(\mathbf{X}_S)\right]}{\sqrt{\mathbb{E}\left[\widehat{f}_S(\mathbf{X}_S)^2\right]}} = 0. \tag{107}$$

As a result, for every $\emptyset \subsetneq T \subsetneq S \subseteq [d]$, the empirical cosine between $\widehat{f}_S$ and $\widehat{f}_T$ coincides with their empirical correlation, since both components are centered. Naturally, *Ortho.* is computed from this quantity.

Remark D.4. Although the theoretical decomposition provably satisfies the constraint (3), exact hierarchical orthogonality is difficult to achieve in practice, especially on finite samples combined with imperfect density estimation.

For the analytical setting, we report in Tab 7 the empirical cosines, in absolute value, between the theoretical ANOVA components. The hierarchical orthogonality constraint is well satisfied in this case: all corresponding cosines are below 10^{-2}, confirming numerically the theoretical result.

Table 7: Empirical cosines between theoretical ANOVA components on the analytical setting, reported as mean ± standard deviation over 100 independent runs with $n = 100\,000$ samples each.

(S, T)	Absolute cosine
$(\{1\}, \emptyset)$	0.0027 ± 0.0018
$(\{2\}, \emptyset)$	0.0026 ± 0.0019
$(\{1,2\}, \emptyset)$	0.0024 ± 0.0020
$(\{1,2\}, \{1\})$	0.0022 ± 0.0017
$(\{1,2\}, \{2\})$	0.0028 ± 0.0024

Explained variance. Across all experiments, the trained models prove to be highly additive, which explains the consistently high R^2 reconstruction values reported in Tab 2 and Tab 5. However, our decomposition does not achieve a perfectly unitary R^2, for three reasons. *First*, the truncated decomposition is restricted to interaction orders $K \in \{1, 2\}$ and therefore cannot capture signal carried by higher-order interactions, whenever the underlying model exhibits them. *Second*, our estimator can still be refined: a more accurate density estimator or a more sophisticated (*e.g.*, penalized) strategy for the regularized least-squares problem would improve the fraction of variance recovered. *Third*, and more fundamentally, recovering exactly 100% of the variance is not a desirable target for an explainability method. Methods that enforce the *efficiency* axiom by construction, such as the SHAP family, distribute the totality of the prediction over feature attributions. This is visible in our experiments: SHAP-based estimates produce substantially noisy curves. By relaxing the efficiency requirement, our approach forgoes explaining the last few percent of variance in exchange for attributions that track the genuine additive structure of the signal rather than its residual noise.

D.8 Additional Analytical Case

We provide a supplementary analytical example in an unbounded-density setting, where Assumption 2.1 no longer holds. We observe that the results of this experiment are keeping relevant. Let consider $\mathbf{X} \in [-1, 1]^3$ with $X_j := \tanh(Z_j)$ for $j \in [3]$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and Σ has unit diagonal and constant off-diagonal $\rho > 0$. The marginal densities are

$$p_j(x) = \frac{1}{\sqrt{2\pi}\,(1 - x^2)} \exp\!\left(-\tfrac{1}{2}\sigma^2(x)\right), \tag{108}$$

$$p_{i,j}(x_i, x_j) = \frac{1}{2\pi\sqrt{1-\rho^2}\,(1-x_i^2)(1-x_j^2)} \times \exp\left(-\frac{\sigma^2(x_i) - 2\rho\,\sigma(x_i)\,\sigma(x_j) + \sigma^2(x_j)}{2(1-\rho^2)}\right), \tag{109}$$

where $\sigma := \operatorname{arctanh}$. The theoretical ANOVA components are defined as in the first analytical case; only the denominators change due to the different densities. We set $\rho = 1/2$ and draw samples by generating $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$ and applying $\tanh$ coordinate-wise.

Fig 5 shows the estimated main effects with $n = 10\,000$ samples: all methods closely track the theoretical components on X_1 and X_2, and correctly assign near-zero contribution to X_3. Fig 6 displays the comparison with our estimator, treating f as a black box. Finally, Tab 8 reports the empirical cosines between ANOVA components, confirming that hierarchical orthogonality holds empirically despite the violation of Assumption 2.1.

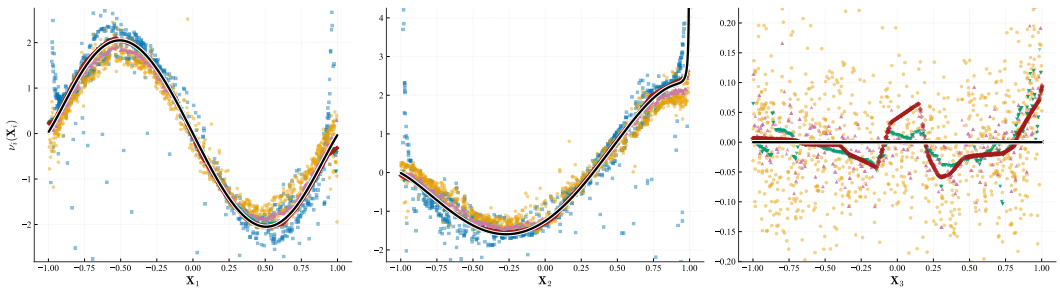


Figure 5: Estimated main effects in the unbounded-density setting for X_1, X_2 and X_3. Comparison with KernelSHAP, TreeSHAP, TreeHFD, NAM and EBM.

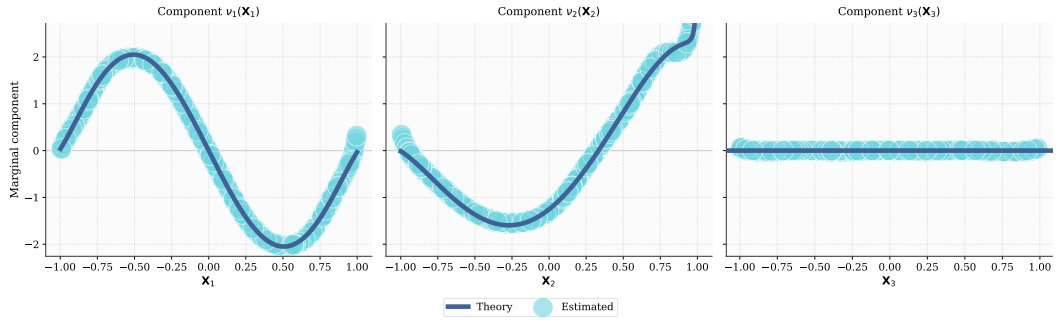


Figure 6: Comparison between the theoretical ANOVA components and our estimator ($K = 2$, $M = 10$, $d_{\text{density}} = 10$, $\varepsilon = 0.01$) in the unbounded-density setting.

Table 8: Empirical cosines between theoretical ANOVA components in the unbounded-density setting (mean $\pm$ std over 100 runs, $n = 100\,000$).

(S, T)	Absolute cosine
$(\{1\}, \emptyset)$	0.0023 ± 0.0017
$(\{2\}, \emptyset)$	0.0025 ± 0.0019
$(\{1, 2\}, \emptyset)$	0.0033 ± 0.0020
$(\{1, 2\}, \{1\})$	0.0017 ± 0.0013
$(\{1, 2\}, \{2\})$	0.0087 ± 0.0132

E Connection with SHAP-based estimators

This section complements the qualitative comparisons of local explanations with global feature-importance results on five datasets: Bike Sharing (**BS**), California Housing (**CH**), Census Income (**CI**), Electrical Grid (**EG**), and Pima Indian Diabetes (**PI**). We compare our method with TreeSHAP for XGBoost models, and with KernelSHAP and DeepSHAP for multilayer perceptrons (MLPs). Each comparison is performed within the same model class. We omit Superconductivity (**SC**) from this additional comparison for computational reasons and to keep the presentation readable.

Evaluation sample. For computational reasons, we compute global importance scores on an evaluation sample of $N_{\text{eval}} = 5\,000$ points for **BS**, **CH**, **CI**, and **EG**. For **PI**, we use all $N_{\text{eval}} = 768$ available observations. For each dataset, all methods use the same evaluation points $(\mathbf{x}^{(i)})_{i=1}^{N_{\text{eval}}}$.

Global importance. For a given method and trained model, let $A \in \mathbb{R}^{N_{\text{eval}} \times d}$ be the matrix of feature contributions on the evaluation sample, with

$$A_{ij} = \begin{cases} \widehat{f}_{\{j\}}(x_j^{(i)}), & \text{the estimated main effect,} \\ \phi_j(\mathbf{x}^{(i)}), & \text{the SHAP-based explainer,} \end{cases} \tag{110}$$

where ϕ_j denotes the SHAP attribution associated with feature j for the corresponding explainer and model. Thus, our scores summarize the estimated main effects. For each feature j, we compute its empirical L_1 magnitude and normalize it across features as follows

$$m_j(A) := \frac{1}{N_{\text{eval}}} \sum_{r=1}^{N_{\text{eval}}} |A_{rj}|, \qquad I_j(A) := 100 \times \frac{m_j(A)}{\sum_{k=1}^{d} m_k(A)}, \tag{111}$$

provided that the denominator is positive. The resulting scores are nonnegative and sum to 100. Tables 9–13 report these normalized mean absolute contributions. The notation X_j identifies the j-th input feature in each dataset.

Rank agreement. For each method, we rank features by their global importance scores $I_j(A)$. Spearman correlation is computed across the d features as the Pearson correlation between the two rank vectors, assigning average ranks to ties. The normalization in Eq. (111) does not change the ranking of the mean absolute contributions.

Table 14 reports these correlations between our method and the corresponding SHAP-based explainer. TreeSHAP is compared with our XGB results, whereas KernelSHAP and DeepSHAP are compared with our MLP results. The final row is the unweighted mean across the five datasets. The mean correlations are 0.9130, 0.9343, and 0.9300, respectively, indicating strong overall agreement in feature rankings. The lowest correlation is 0.7619, obtained on **BS** for the comparison with TreeSHAP. These results quantify agreement in global rankings; they do not establish equivalence of local attributions or accuracy against a ground-truth explanation.

Table 9: Global feature importance (%) on Bike Sharing (**BS**).

	XGBoost		MLP		
Feature	Ours	TreeSHAP	Ours	KernelSHAP	DeepSHAP
X_1	50.3744	45.5180	50.5807	47.4075	53.9525
X_2	1.5195	9.9017	2.1110	10.2345	8.6640
X_3	1.0339	0.6807	0.2932	0.7667	0.7717
X_4	10.7945	9.2791	7.9111	7.5666	7.2755
X_5	20.3323	20.2682	22.7251	20.5042	17.0642
X_6	10.3583	8.8610	12.0030	7.8736	7.2689
X_7	1.8881	2.3650	1.2493	1.8932	1.6535
X_8	3.6991	3.1264	3.1264	3.7538	3.3497

Table 10: Global feature importance (%) on California Housing (**CH**).

	XGBoost		MLP		
Feature	Ours	TreeSHAP	Ours	KernelSHAP	DeepSHAP
X_1	33.2500	29.2700	21.8700	19.4600	11.3300
X_2	4.8300	4.2100	3.5700	2.9700	1.6400
X_3	5.4900	7.3600	2.9400	9.8500	3.6800
X_4	4.7700	1.7100	2.4900	2.0600	1.3800
X_5	1.4400	1.4700	1.1200	1.4900	0.7300
X_6	14.6200	13.5500	10.6000	12.4700	6.5800
X_7	17.5400	23.0300	30.9100	28.6800	39.1100
X_8	18.0600	19.3800	26.5000	23.0100	35.5600

Table 11: Global feature importance (%) on Census Income (**CI**).

	XGBoost		MLP		
Feature	Ours	TreeSHAP	Ours	KernelSHAP	DeepSHAP
X_1	13.3054	15.2164	14.9586	17.0881	17.0826
X_2	6.8988	1.4843	1.3829	1.3580	1.9452
X_3	1.1135	1.6579	0.8368	1.0403	1.5745
X_4	4.8997	0.9003	8.9546	4.7806	4.6948
X_5	12.0810	13.5861	13.6308	14.8325	12.4089
X_6	18.5472	18.5625	19.8913	16.5109	16.6904
X_7	10.1375	7.0298	2.5592	2.8031	3.1396
X_8	8.4870	14.1087	4.3472	8.7373	9.1030
X_9	0.7372	0.5729	1.4692	1.6415	1.8181
X_{10}	4.8173	2.6871	7.5560	7.2096	6.1510
X_{11}	8.9345	11.8062	12.6143	10.0592	10.5772
X_{12}	2.5121	3.8328	1.9152	1.7506	2.6594
X_{13}	7.1106	8.0276	9.4338	11.7770	11.5470
X_{14}	0.4182	0.5275	0.4502	0.4113	0.6083

Table 12: Global feature importance (%) on Electrical Grid (**EG**).

	XGBoost		MLP		
Feature	Ours	TreeSHAP	Ours	KernelSHAP	DeepSHAP
X_1	13.6437	13.4119	13.2043	12.1318	11.5254
X_2	13.7234	13.4129	13.1554	12.2072	11.8409
X_3	13.5724	14.0789	13.6184	13.8353	14.0872
X_4	13.5033	12.7727	13.0518	12.9702	13.0222
X_5	0.1487	0.4979	0.0882	1.1253	1.5344
X_6	0.3914	0.5186	0.4190	0.7438	1.2605
X_7	0.4831	0.5082	0.1860	0.7719	1.1540
X_8	0.6050	0.6157	0.2197	0.6337	1.0468
X_9	10.4042	10.5454	11.0589	11.1615	11.2422
X_{10}	11.4173	11.3722	11.7994	11.7145	11.4116
X_{11}	11.3252	11.6519	11.7348	11.1255	10.6959
X_{12}	10.7824	10.6137	11.4641	11.5791	11.1788

Table 13: Global feature importance (%) on Pima Indian Diabetes (**PI**).

	XGBoost		MLP		
Feature	Ours	TreeSHAP	Ours	KernelSHAP	DeepSHAP
X_1	5.5059	4.9248	12.2998	12.9708	12.8868
X_2	35.2554	33.4458	36.5302	36.1625	34.8230
X_3	3.1025	3.6014	7.4437	7.7228	8.1034
X_4	1.8986	4.0152	1.9596	1.6997	2.2543
X_5	4.6394	4.8442	1.8499	2.1823	2.4806
X_6	19.3457	18.7768	20.4796	20.7625	20.8446
X_7	12.7497	12.6906	9.7095	9.1647	9.1810
X_8	17.5028	17.7012	9.7277	9.3346	9.4262

Table 14: Spearman rank correlations between our global feature importances and those of SHAP-based explainers. Comparisons use XGBoost for TreeSHAP and MLPs for KernelSHAP and DeepSHAP. The mean is computed with equal weight for each dataset.

	XGBoost	MLP	
Dataset	vs. TreeSHAP	vs. KernelSHAP	vs. DeepSHAP
BS	0.7619	0.8571	0.8333
CH	0.9762	0.9762	0.9762
CI	0.8857	0.9736	0.9692
EG	0.9650	0.8881	0.8951
PI	0.9762	0.9762	0.9762
Mean	**0.9130**	**0.9343**	**0.9300**

F Additional Figures

F.1 Bike Sharing Dataset

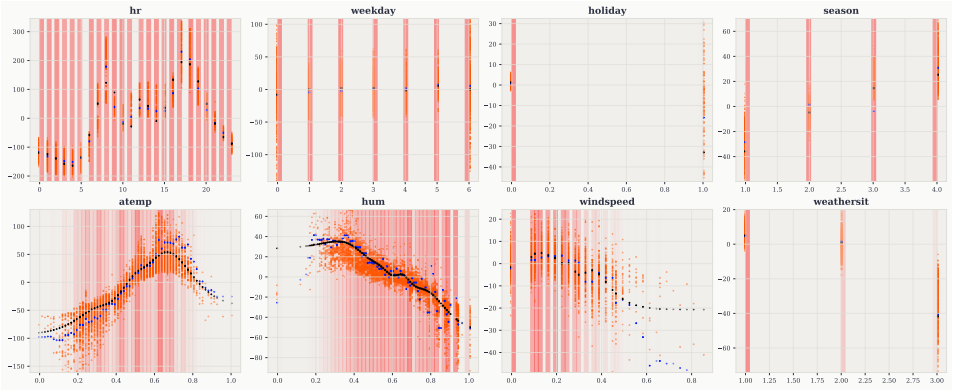


Figure 7: Estimated main effects on *Bike Sharing*: our method (black) vs TreeHFD (main effects) and TreeSHAP on a trained XGB.

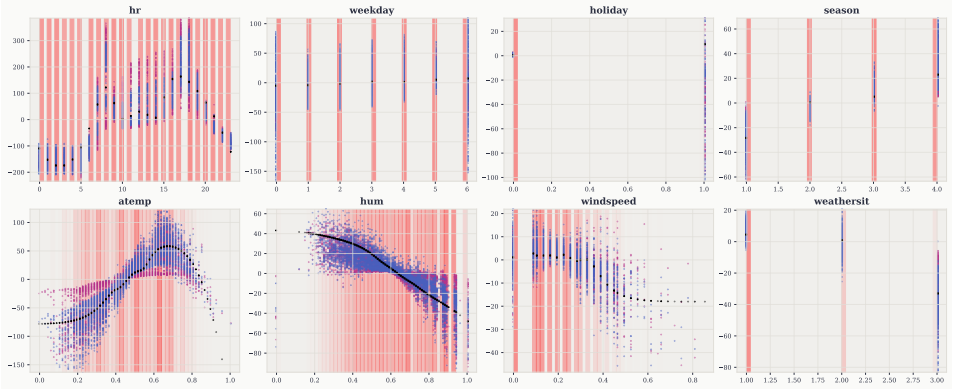


Figure 8: Estimated main effects on *Bike Sharing*: our method (black) vs KernelSHAP and DeepSHAP on a trained MLP.

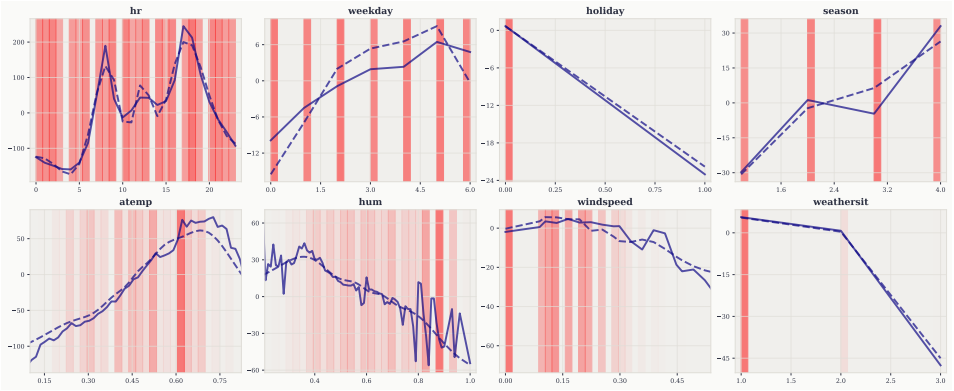


Figure 9: Comparison of native main effects from an EBM with those recovered by our method on *Bike Sharing*. EBM (solid) vs our method (dashed).

F.2 California Housing Dataset

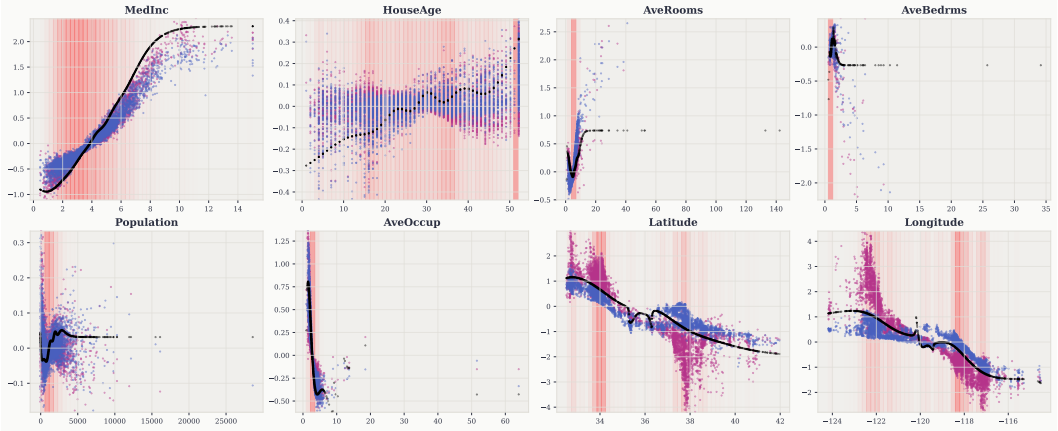


Figure 10: Estimated main effects on *California Housing*: our method (black) vs KernelSHAP and DeepSHAP on a trained MLP.

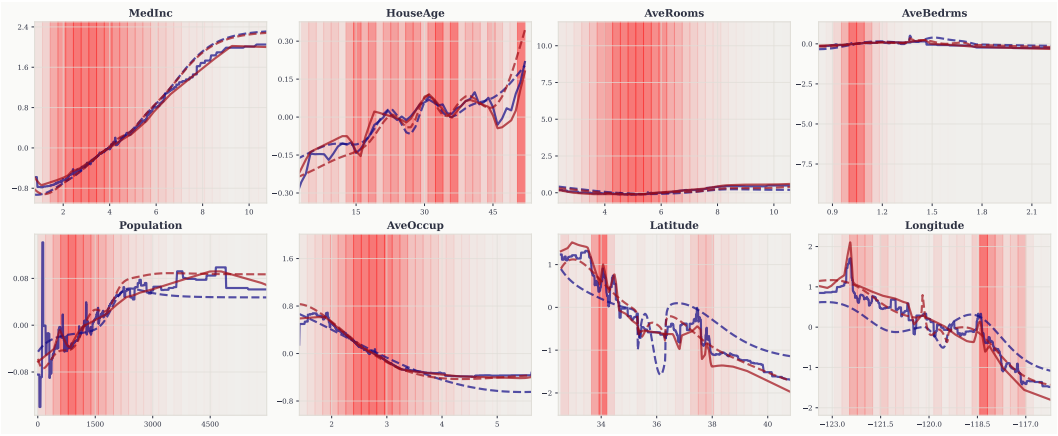


Figure 11: Comparison of native main effects from an EBM with those recovered by our method on *California Housing*. EBM (solid) vs our method (dashed); NAM (solid) vs our method (dashed).

F.3 CENSUS INCOME DATASET

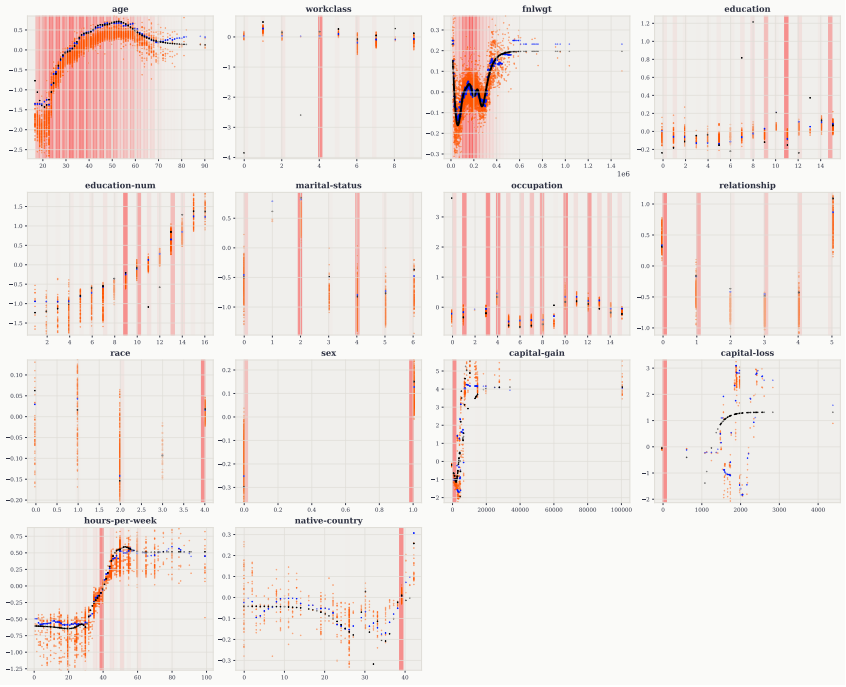


Figure 12: Estimated main effects on *Census Income*: our method (black) vs TreeHFD (main effects) and TreeSHAP on a trained XGB.

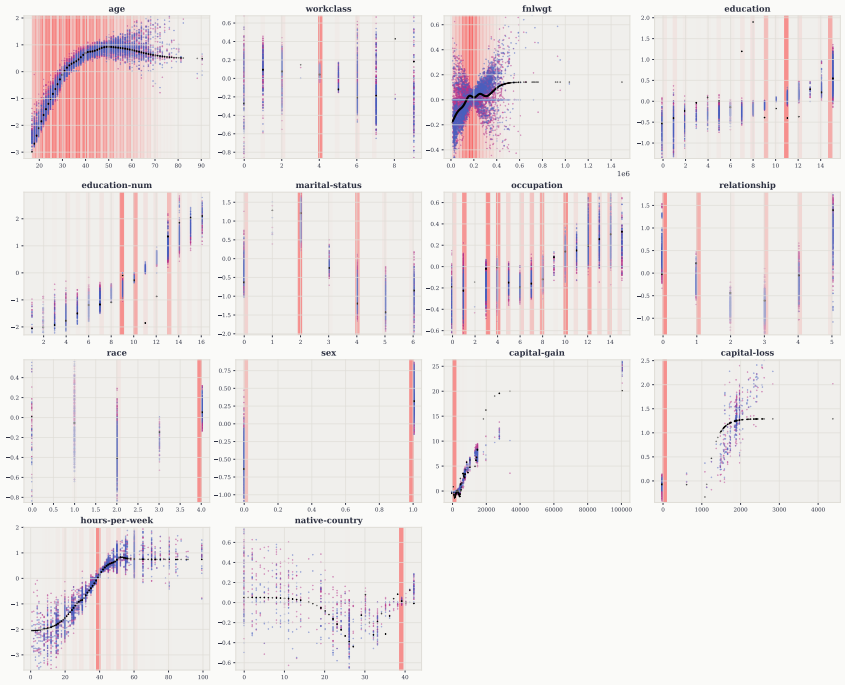


Figure 13: Estimated main effects on *Census Income*: our method (black) vs KernelSHAP and DeepSHAP on a trained MLP.

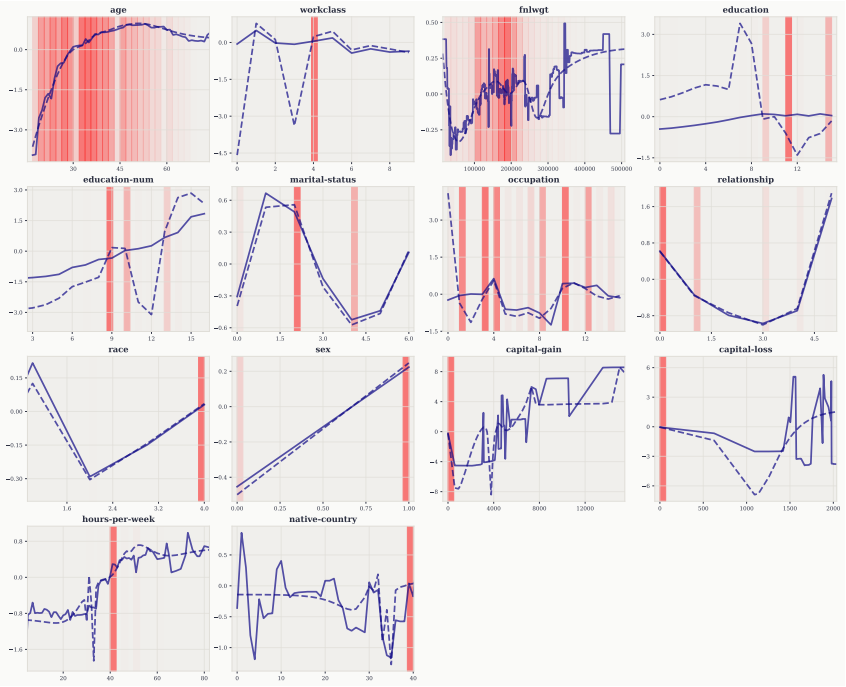


Figure 14: Comparison of native main effects from an EBM with those recovered by our method on *Census Income*. EBM (solid) vs our method (dashed).

F.4 Electrical Grid Dataset

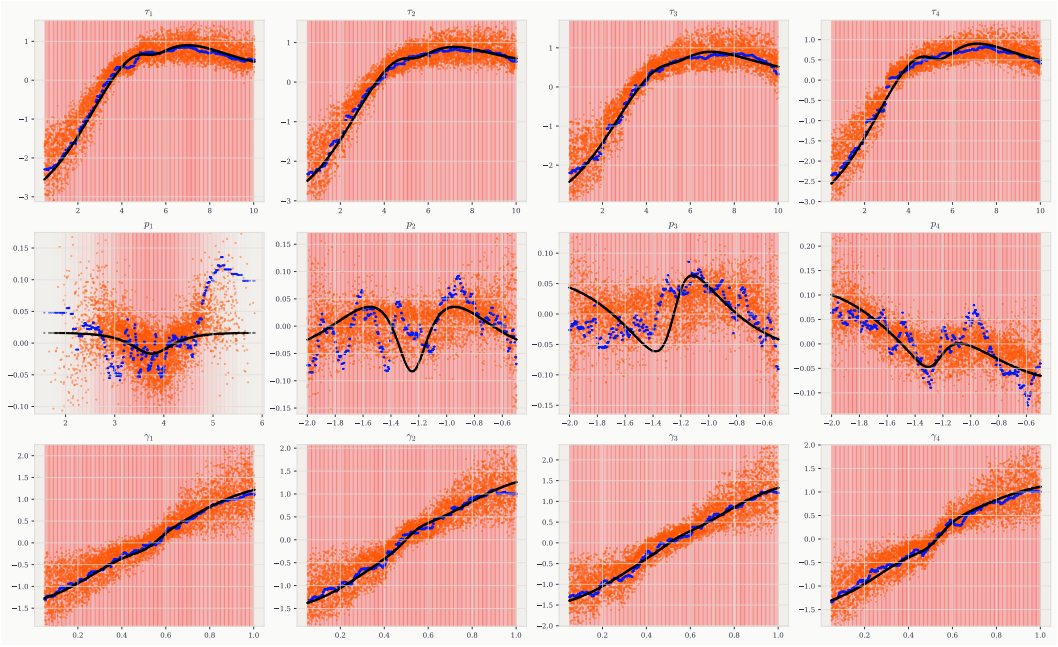


Figure 15: Estimated main effects on *Electrical Grid*: our method (black) vs TreeHFD (main effects) and TreeSHAP on a trained XGB.

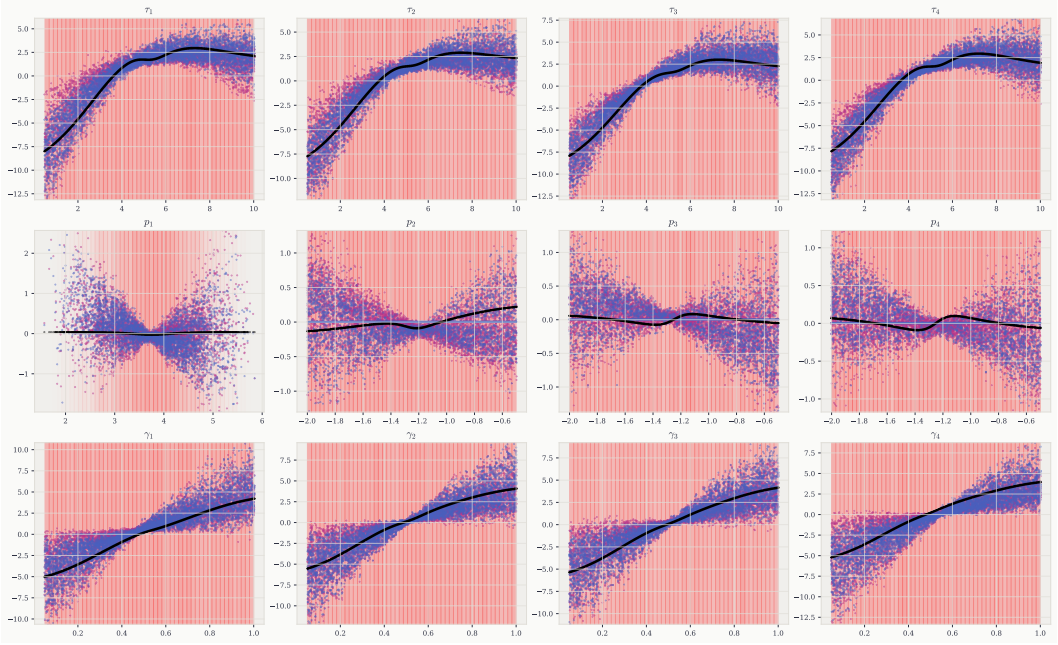


Figure 16: Estimated main effects on *Electrical Grid*: our method (black) vs KernelSHAP and DeepSHAP on a trained MLP.

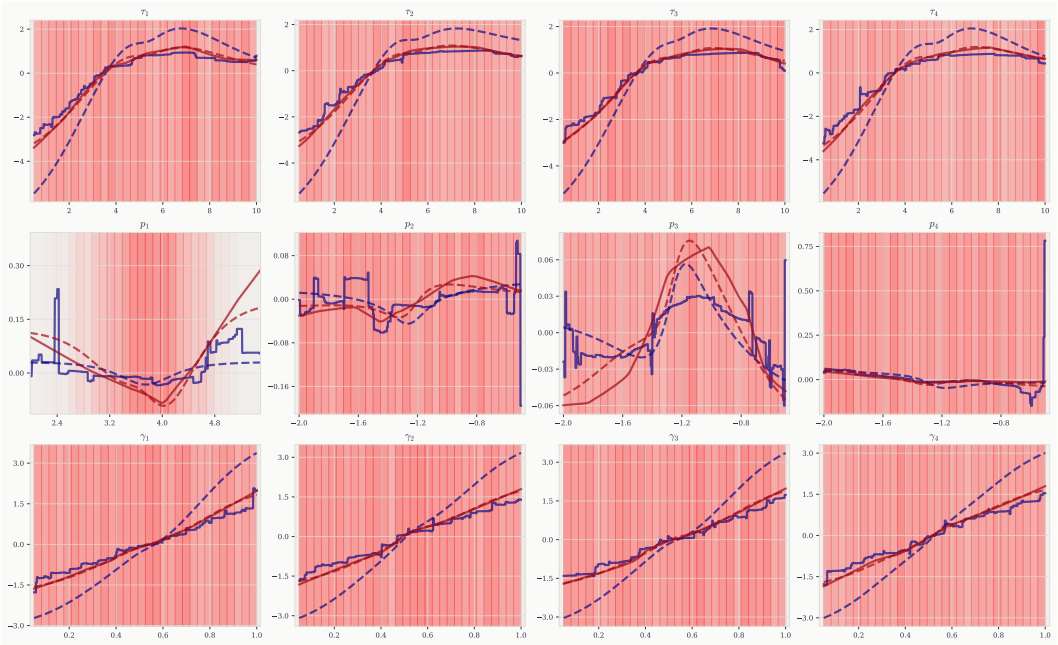


Figure 17: Comparison of native main effects from an EBM with those recovered by our method on *Electrical Grid*. EBM (solid) vs our method (dashed); NAM (solid) vs our method (dashed).

F.5 DIABETES DATASET

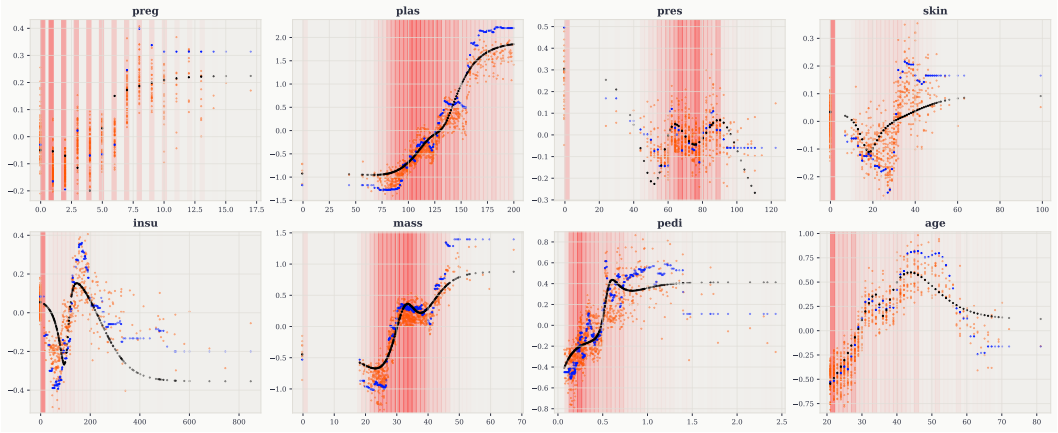


Figure 18: Estimated main effects on *Diabetes*: our method (black) vs TreeHFD (main effects) and TreeSHAP on a trained XGB.

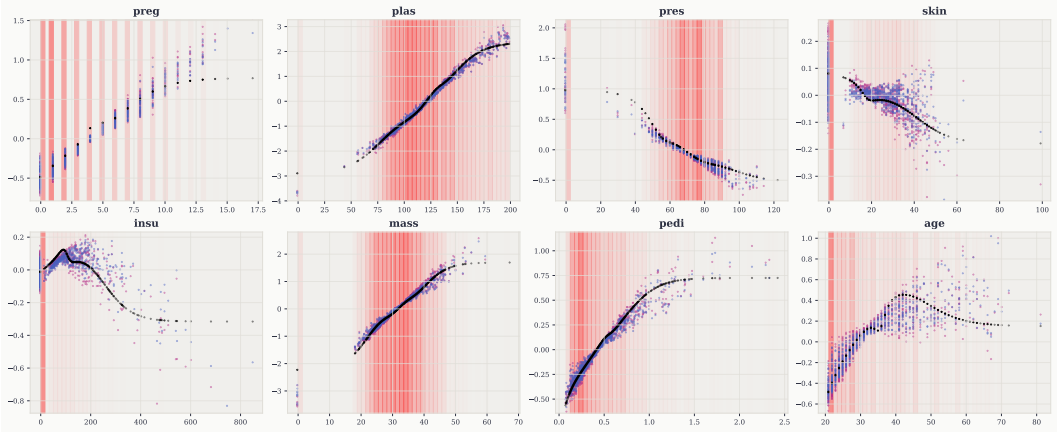


Figure 19: Estimated main effects on *Diabetes*: our method (black) vs KernelSHAP and DeepSHAP on a trained MLP.