Bachelor's Thesis

Functional ANOVA Decomposition

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, July 31th, 2025

Submitted in partial fulfillment of the requirements for the degree of B. Sc.
Supervised by Prof. Dr. Thomas Nagler

Abstract

This article studies the functional ANOVA decomposition (fANOVA) in the context of model interpretability. We begin by introducing the classical fANOVA, which assumes independent inputs, and illustrate its equivalence to the Hoeffding decomposition under zero-centered variables with an example. We then unify different notations under the concept of orthogonal projections and briefly present the variance decomposition. Next, we extend fANOVA to settings with dependent inputs, discussing two different formalizations and highlighting why one is more suitable for deriving an explicit solution in an exemplary decomposition, while the other remains primarily theoretical. Finally, we adopt an applied perspective, visualizing the decomposition of various functions and providing a conceptual overview of current estimation approaches.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Thomas Nagler, whose guidance and support were essential to the success of this thesis. His expertise and constructive feedback not only shaped the direction of this work but also made the process of writing it both enjoyable and rewarding. I am also thankful for the presence and encouragement of my friends during this period, which provided me with motivation and balance alongside my studies.

Contents

1 Introduction **1**

2 Background and Related Work **3**

3 Formalization of fANOVA **5**

3.1 Classical fANOVA 6

3.1.1 Construction of the fANOVA Component Functions 8

3.1.2 Example: Independent Multivariate Normal Inputs 9

3.1.3 Equality to Hoeffding Decomposition 10

3.1.4 fANOVA via Projection 13

3.1.5 Variance Decomposition 15

3.2 Generalized fANOVA 18

3.2.1 Conditions for Generalized fANOVA 19

3.2.2 Construction of the Generalized fANOVA Component Functions . . 22

3.2.3 Generalization via Projection 27

3.2.4 Generalized Variance Decomposition 29

3.2.5 Example: Dependent Multivariate Normal Inputs 29

4 Visualization and Estimation **34**

4.1 Comparison of Decompositions 34

4.2 Comparison of Functions 36

4.2.1 Linear 36

4.2.2 Linear and Quadratic 36

4.2.3 Interaction 37

4.2.4 Full 39

4.3 Estimation of fANOVA Component Functions 41

5 Conclusion **44**

References **44**

A Appendix **V**

B Electronic appendix **VIII**

List of Figures

1 Main fANOVA component functions (left) and interaction component (right) of $h(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with independent inputs. 34

2 Main effects (left) and interaction effect (right) from a fANOVA-type decomposition of $h(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with dependent inputs, $\rho = 0.5$. 35

3 Main fANOVA component functions (left) and interaction component (right) from the generalized fANOVA decomposition of $h(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with dependent inputs, $\rho = 0.5$. 36

4 Main fANOVA component functions $q_{\{1\}}(x_1) = a_1x_1$ and $q_{\{2\}}(x_2) = a_2x_2$ of the linear function $q(x_1, x_2) = a_1x_1 + a_2x_2$. 37

5 Main fANOVA component functions $p_{\{1\}}(x_1) = a_1x_1 + a_{11}(x_1^2 - 1)$ and $p_{\{2\}}(x_2) = a_2x_2 + a_{22}(x_2^2 - 1)$ of the polynomial $p(x_1, x_2) = a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{22}x_2^2$. 38

6 Main fANOVA component functions (left) and interaction component (right) of the function $w(x_1, x_2) = 2x_1x_2$ for varying ρ. 40

7 Main fANOVA component functions (left) and interaction component (right) of the polynomial $z(x_1, x_2) = 2x_1 + x_1^2 + 0.5x_1x_2$ for varying ρ. The coefficient sets are identical while the correlation structure varies. 41

8 Main fANOVA component functions (left) and interaction component (right) of the polynomial $z(x_1, x_2) = -2x_1 - 2x_2 + x_1^2 - x_2^2 + x_1x_2$ for varying ρ. The coefficient sets are identical, while the correlation structure varies. . . 42

1 Introduction

With the rise of machine learning (ML) and increasingly more complex probabilistic models, interpretability has become a major concern for practitioners and researchers alike. One of the foundational mathematical methods supporting the goal of interpretability is the functional ANOVA decomposition (fANOVA).

At its core, the fANOVA decomposition provides a method which allows decomposing integrable functions into the sum of mutually orthogonal component functions of varying dimensionality. It is not only useful in interpretability of black box models (Hooker (2004), Molnar (2025)), but also in areas, such as uncertainty quantification of complex systems (Rahman, 2014), non-parametric statistical modelling (see for example Stone et al. (1997)), sensitivity analysis (Sobol, 1993), and many more fields. Given this wide range of applications, fANOVA is an essential concept worth understanding in depth.

However, learning about fANOVA is not straightforward. A problem is the mix of formalizations and definitions around the method, partly due to its long history and the various streams of science that have used it. This already starts with the name of the method. It has been called decomposition into summands of different order (Sobol, 1993), ANOVA representation (Sobol, 2001), functional ANOVA decomposition (Hooker, 2004), ANOVA dimensional decomposition (Rahman, 2014), or Hoeffding-Sobol' decomposition (Chastaing et al., 2012) – in this thesis we will refer to it as the fANOVA decomposition.

The variety does not stop at naming; authors also differ in how they formalize the decomposition, often using distinct notation, slightly different sets of assumptions, and either interpret fANOVA from a probabilistic perspective, using expectations, or from a more deterministic mathematical viewpoint, using integrals. While these approaches are mathematically equivalent and can be unified under the concept of orthogonal projections, this connection is often not obvious when first encountering the literature.

Given this state of affairs, there is a clear need for a comprehensive overview of fANOVA-related work and for a unification of the various notations and definitions that ultimately express the same concepts. Bringing clarity into the fANOVA landscape is more relevant than ever as the method has recently attracted renewed attention in interpretable machine learning (IML) literature (see for example Hu et al. (2025)), yet the theoretical foundation is often mentioned only briefly or left implicit.

This thesis addresses that gap by providing an accessible and intuitive introduction to the fANOVA decomposition while remaining mathematically rigorous. It can be viewed as a handbook of the fANOVA decomposition that will help researchers and practitioners to understand the mathematical background of this method as well as its more applied aspects.

This work is organized as follows: It starts with background and related work (section 2). This is followed by the central part in which we give the formal definition of the classical and generalized fANOVA decomposition (section 3). Next, we illustrate the characteristics of the method using analytical examples. We then briefly outline current estimation schemes (section 4) and conclude with a discussion and possible future research directions.

2 Background and Related Work

The literature around fANOVA can be grouped into several thematic clusters. Each highlights a different angle on why fANOVA has proven useful and points to why a unified presentation is needed.

The underlying principle of the hierarchical, additive decomposition of a function dates back to Hoeffding (1948). In his seminal work on U-statistics, he introduced the Hoeffding decomposition. Though originally framed around estimators, this decomposition laid the groundwork for fANOVA by showing how a symmetric function can be written as a sum of mutually orthogonal component functions of increasing dimensionality.

Independently, Sobol (1993) proved that any square integrable function on the unit hypercube can be decomposed into a sum of mutually orthogonal and zero-centered component functions. The foundational work on fANOVA shows, that it is rooted in rigorous mathematical theory, and provides a principled way to break down complex multivariate functions into interpretable, orthogonal parts.

A second strand of work explores how fANOVA underlies non-parametric modeling. Takemura (1983) introduced tensor-analysis of ANOVA decompositions, laying the theoretical foundation. Stone (1994) applied fANOVA ideas to polynomial splines and generalized additive models. Gu (2013) extended this into smoothing-spline ANOVA frameworks for flexible regression estimation. Their work demonstrates, that fANOVA not only provides a theoretical decomposition, but also serves as a basis for widely-used non-parametric statistical models featuring additive structure and controlled interactions.

Perhaps the most well-known application of fANOVA is in variance-based sensitivity analysis. Sobol's original decomposition led directly to a variance decomposition, on which Sobol' indices are based. Work from Owen (2013, 2014) modernized this framework, introducing efficient estimation strategies and generalized indices suited to quasi-Monte Carlo methods. Borgonovo et al. (2022) further advanced the field with mixture-based generalizations of fANOVA for uncertainty quantification.

Classical fANOVA requires independent input variables, which is a strong assumption in many real-world applications. Therefore, a stream of literature is concerned with the generalization of fANOVA to dependent variables. While Hooker (2007) was the first to present a generalized fANOVA framework, many other researchers were inspired by his work to create modifications of this (Rahman, 2014, Chastaing et al., 2012, Il Idrissi et al., 2025). We see the generalization as central part of the basis of the fANOVA decomposition and therefore will also present it in this thesis.

A recent cluster of literature studies fANOVA for model interpretability. There is work of Lengerich et al. (2020), König et al. (2024), Choi et al. (2025) that all enhance interpretability by using fANOVA to identify and disentangle variable interactions. Then there is work done in the explicit context of IML, where fANOVA can be used as a model-agnostic tool (Hooker, 2004, Fumagalli et al., 2025) or as foundational principle to build inherently interpretable models (Hu et al., 2025). fANOVA-based interpretability methods is probably the most novel field of fANOVA in which research is actively ongoing.

Finally, there are specific domains of statistics, such as geostatistics, where fANOVA-based Kriging models are designed (Muehlenstaedt et al., 2012) or complex functions arising in computational finance are studied (Liu and Owen, 2006).

3 Formalization of fANOVA

The formal setup is based on Rahman (2014) and Chastaing et al. (2012). Let $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$, and $\mathbb{R}_0^+$ denote the sets of positive integer (natural), nonnegative integer, real, and nonnegative real numbers, respectively. Throughout this thesis, we represent the k-dimensional Euclidean space by $\mathbb{R}^k$ and the set of all $k \times k$ real-valued matrices by $\mathbb{R}^{k\times k}$.

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space, where Ω is a sample space, $\mathcal{F}$ is a $\sigma-$algebra on Ω and $\nu : \mathcal{F} \to [0, 1]$ is a probability measure. $\mathcal{B}^N$ is the Borel σ-algebra on $\mathbb{R}^N, N \in \mathbb{N}$. $\boldsymbol{X} = (X_1, \ldots, X_N) : (\Omega, \mathcal{F}) \to (\mathbb{R}^N, \mathcal{B}^N)$ denotes an $\mathbb{R}^N$-valued random vector. We assume that the probability distribution of $\boldsymbol{X}$ is continuous and completely defined by the joint probability density function $f_{\boldsymbol{X}} : \mathbb{R}^N \to \mathbb{R}_0^+$.

Let u denote a subset of $\{1, \ldots, N\}$ with the complementary set $-u := \{1, \ldots, N\}\backslash u$ and cardinality $0 \le |u| \le N$. We denote strict inclusion of a subset by $\subsetneq$ and $\subseteq$ allows for equality. $\boldsymbol{X_u} = (X_{i_1}, \ldots, X_{i_{|u|}})$, $u \ne \emptyset$, $1 \le i_1 < \cdots < i_{|u|} \le N$ is a subvector of $\boldsymbol{X}$ and $\boldsymbol{X}_{-u} = \boldsymbol{X}_{\{1,\ldots,N\}\backslash u}$ is the complementary subvector.

The marginal density function of $\boldsymbol{X_u}$ is $f_u(\boldsymbol{x_u}) := \int_{\mathbb{R}^{N-|u|}} f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-u})$ for a given set $\emptyset \ne u \subseteq \{1, \ldots N\}$.

Let $y(\boldsymbol{X}) := y(X_1, \ldots, X_N)$ be a real-valued, measurable transformation on $(\Omega, \mathcal{F})$, which represents a probabilistic model with random variables as inputs. The Hilbert space of square-integrable functions y with respect to the induced generic measure $f_{\boldsymbol{X}}(\boldsymbol{x})d\nu(\boldsymbol{x})$ supported on $\mathbb{R}^N$ is given by:

$$\mathcal{L}^2(\Omega, \mathcal{F}, \nu) = \left\{ y : \Omega \to \mathbb{R} \ s.t. \ \mathbb{E}[y^2(\boldsymbol{X})] < \infty \right\}.$$

The inner product is defined by:

$$\langle y, g \rangle = \int_{\mathbb{R}^N} y(\boldsymbol{x})g(\boldsymbol{x}) \, f_{\boldsymbol{X}}(\boldsymbol{x})d\nu(\boldsymbol{x}) = \mathbb{E}[y(\boldsymbol{X})g(\boldsymbol{X})], \quad \forall y, g \in \mathcal{L}^2.$$

The norm, denoted as $\|.\|$, is defined by:

$$\|y\| = \sqrt{\langle y, y \rangle} = \sqrt{\int_{\mathbb{R}^N} y^2(\boldsymbol{x}) \, d\nu(\boldsymbol{x})} = \mathbb{E}[y^2(\boldsymbol{X})], \quad \forall y \in \mathcal{L}^2.$$

We start by defining the fANOVA decomposition in a general form, which is independent of distribution assumptions about the input variables or anything of the sort. Its specific form is determined by the assumptions about the input variables and integration measure.

Definition 3.1. *Let y denote a mathematical model with input vector $\boldsymbol{X} := (X_1, \ldots, X_N)$. The functional ANOVA (fANOVA) decomposition of y takes the form:*

$$y(\boldsymbol{X}) = \sum_{u \subseteq \{1,\ldots,N\}} y_u(\boldsymbol{X}_u). \tag{1}$$

The functions y_u are referred to as fANOVA component functions *(or simply* components*) throughout this thesis.*

3.1 Classical fANOVA

For his original fANOVA decomposition, Sobol' only considered functions defined on the unit hypercube, but later work shows that it is no problem to work within the measure space $(\mathbb{R}^N, \mathcal{B}^N, \nu)$. In any case, we assume that the coordinates $X_1, \ldots, X_N$ are independent of each other. Under independence, we work with a product-type probability measure of $\boldsymbol{X}$ given by $f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) = \prod_{i=1}^N f_{\{i\}}(x_i)\, d\nu(x_i)$, where $f_{\{i\}} : \mathbb{R} \to \mathbb{R}_0^+$ is the marginal probability density function of X_i defined on $(\Omega_i, \mathcal{F}_i, \nu_i)$ with a bounded or an unbounded support on $\mathbb{R}$.

Given this setup, we formulate a condition, proposed by Rahman (2014), which ensure that the fANOVA component functions are well-defined and interpretable.

Condition 3.1 (Strong annihilating conditions, Rahman (2014))**.** *For the classical fANOVA decomposition we require, that all the nonconstant fANOVA component functions y_u integrate to zero w.r.t. the marginal probability density of each random variable in u, that is,*

$$\int_{\mathbb{R}} y_u(\boldsymbol{x}_u) f_{\{i\}}(x_i)\, d\nu(x_i) = 0 \quad \textit{for } i \in u \neq \emptyset. \tag{2}$$

Proposition 3.1 (Rahman (2014))**.** *Given that the strong annihilating conditions are satisfied, the fANOVA component functions y_u, where $\emptyset \neq u \subseteq \{1, \ldots, N\}$, are centered around zero:*

$$\int_{\mathbb{R}^N} y_u(\boldsymbol{x}_u)\, f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) = \mathbb{E}[y_u(\boldsymbol{X}_u)] = 0. \tag{3}$$

Proof. Given an index $i \in u$, we write:

$$\begin{aligned}\mathbb{E}[y_u(\boldsymbol{X}_u)] &:= \int_{\mathbb{R}^N} y_u(\boldsymbol{x_u}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\ &= \int_{\mathbb{R}^{|u|}} y_u(\boldsymbol{x_u}) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u)\, d\nu(\boldsymbol{x}_u) \\ &= \int_{\mathbb{R}^{|u|}} y_u(\boldsymbol{x_u}) \prod_{j\in u} f_{\{j\}}(x_j)\, d\nu(\boldsymbol{x}_u) \\ &= \int_{\mathbb{R}^{|u|-1}} \int_{\mathbb{R}} y_u(\boldsymbol{x_u}) f_{\{i\}}(x_i)\, d\nu(x_i) \prod_{j\in u, j\neq i} f_{\{j\}}(x_j)\, d\nu(x_{u\setminus\{i\}}) = 0.\end{aligned}$$

□

Proposition 3.2 (Rahman (2014))**.** *Given the strong annihilating conditions are satisfied, two distinct fANOVA component functions y_u and y_v, where $\emptyset \neq u \subseteq \{1, \ldots, N\}$, $\emptyset \neq v \subseteq \{1, \ldots, N\}$, and $u \neq v$, are orthogonal; i.e., they satisfy*

$$\int_{\mathbb{R}^N} y_u(\boldsymbol{x_u}) y_v(\boldsymbol{x_v}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \mathbb{E}[y_u(\boldsymbol{X}_u) y_v(\boldsymbol{X}_v)] = 0. \tag{4}$$

Proof. Since $u \neq v$, there exists at least one index contained in exactly one of the sets. Without loss of generality, we pick $i \in u \setminus v$. Then $y_v(\boldsymbol{x_v})$ is independent of x_i, and assuming the strong annihilating conditions hold, we have:

$$\int_{\mathbb{R}} y_u(\boldsymbol{x}_u) f_{\{i\}}(x_i)\, d\nu(x_i) = 0 \quad \text{for all fixed } \boldsymbol{x}_{u\setminus\{i\}}.$$

Hence,

$$\begin{aligned}\mathbb{E}[y_u(\boldsymbol{X}_u) y_v(\boldsymbol{X}_v)] &= \int_{\mathbb{R}^N} y_u(\boldsymbol{x_u}) y_v(\boldsymbol{x_v}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\ &= \int_{\mathbb{R}^N} y_u(\boldsymbol{x}_u) y_v(\boldsymbol{x}_v) \prod_{j=1}^{N} f_{\{j\}}(x_j)\, d\nu(\boldsymbol{x}) \\ &= \int_{\mathbb{R}^{N-1}} \left(\int_{\mathbb{R}} y_u(\boldsymbol{x}_u) f_{\{i\}}(x_i)\, d\nu(x_i) \right) y_v(\boldsymbol{x}_v) \prod_{j\neq i} f_{\{j\}}(x_j)\, d\nu(\boldsymbol{x}_{-i}) \\ &= 0.\end{aligned}$$

□

As we have seen, the fANOVA component functions are "fully orthogonal" to each other, meaning not only components of different order are orthogonal to each other but also ones of the same order are. These properties are desirable because they ensure that the components can be interpreted as isolated effects of specific variables or their interactions. For example, the component function $y_{\{1\}}$ represents the isolated main effect of X_1; no other contributions involving X_1 through interactions with other variables are mixed into it. Similarly, the component function $y_{\{1,2\}}$ captures only the interaction effect between X_1 and X_2, while the individual effect of X_1 is already represented by $y_{\{1\}}$ and therefore does not merge into $y_{\{1,2\}}$. From the perspective of interpretability, this clean separation of effects distinguishes the fANOVA decomposition from alternative methods such as partial dependence (PD) or Shapley values.

3.1.1 Construction of the fANOVA Component Functions

The individual fANOVA component functions associated with the variables indexed by u are obtained by integrating the original function $y(\boldsymbol{X})$ with respect to all variables except those in u and subtracting the corresponding lower-order components. Intuitively, the integration averages out the influence of all other variables, leaving a function of the variables of interest only. Subtracting the lower-order components removes effects already explained by other variables or interactions, yielding the isolated effects of the variables in u.

For the classical fANOVA decomposition, these components can be computed as described in Rahman (2014). The constant component (with $u = \emptyset$), is obtained by integrating over all variables:

$$y_\emptyset = \int_{\mathbb{R}^N} y(\boldsymbol{x}) \prod_{i=1}^{N} f_{\{i\}}(x_i)\, d\nu(x_i) = \mathbb{E}[y(\boldsymbol{X})]. \tag{5}$$

For all other components, where $\emptyset \neq u \subseteq \{1, \ldots, N\}$, we obtain:

$$y_u(\boldsymbol{X}_u) = \int_{\mathbb{R}^{N-|u|}} y(\boldsymbol{X}_u, \boldsymbol{x}_{-u}) \prod_{i=1, i\notin u}^{N} f_{\{i\}}(x_i)\, d\nu(x_i) - \sum_{v \subsetneq u} y_v(\boldsymbol{X}_v). \tag{6}$$

Notice that this definition relies on a product-type measure rooted in the independence assumption. We will see what changes when we let go of this assumption in the second part of this section. As suggested earlier, the fANOVA component functions offer a clear interpretation of the model, decomposing it into main effects, two-way interaction effects, and so on. This is why fANOVA decomposition has received increasing attention in the IML literature, holding the potential for a global model-agnostic explanation method of black box models.

3.1.2 Example: Independent Multivariate Normal Inputs

Throughout this thesis, we use the following simple setup as a running example.

Example 3.1 (Running Example). *Consider the bivariate function*

$$h(x_1, x_2) = a + x_1 + 2x_2 + x_1 x_2, \tag{7}$$

which includes both main effects and an interaction term.

Assume the input vector

$$\boldsymbol{X} = (X_1, X_2)^\mathsf{T}$$

follows a bivariate standard normal distribution

$$\boldsymbol{X} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

covering both independent inputs ($\rho = 0$) and correlated inputs ($\rho \neq 0$).

From properties of the multivariate normal distribution, the marginal distributions are

$$X_1 \sim \mathcal{N}(0, 1), \qquad X_2 \sim \mathcal{N}(0, 1),$$

and the conditional distributions are given by:

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}(\rho x_2, 1 - \rho^2), \quad X_2 \mid X_1 = x_1 \sim \mathcal{N}(\rho x_1, 1 - \rho^2).$$

The classical fANOVA decomposition we covered so far assumes independence, i.e., $\rho = 0$. Here, X_1 and X_2 are independent and standard normal, so the conditional means vanish, and the classical fANOVA decomposition simplifies considerably. Computing the constant component via expectation yields:

$$\begin{aligned} h_\emptyset &= \mathbb{E}[h(X_1, X_2)] \\ &= \mathbb{E}[a + X_1 + 2X_2 + X_1 X_2] \\ &= a + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1]\mathbb{E}[X_2] = a. \end{aligned}$$

Given the zero-mean property and independence, the main components and the interaction components can be computed as follows:

$$
\begin{aligned}
h_{\{1\}}(x_1) &= \mathbb{E}_{X_2}[h(x_1, X_2)] - h_\emptyset \\
&= \mathbb{E}_{X_2}[a + x_1 + 2X_2 + x_1X_2] - a \\
&= x_1 + 2\mathbb{E}_{X_2}[X_2] + x_1\mathbb{E}_{X_2}[X_2] = x_1, \\
h_{\{2\}}(x_2) &= \mathbb{E}_{X_1}[h(X_1, x_2)] - h_\emptyset \\
&= \mathbb{E}_{X_1}[a + X_1 + 2x_2 + X_1x_2] - a \\
&= \mathbb{E}_{X_1}[X_1] + 2x_2 + x_2\mathbb{E}_{X_1}[X_1] = 2x_2, \\
h_{\{1,2\}}(x_1, x_2) &= \mathbb{E}[h(x_1, x_2)] - h_\emptyset - h_{\{1\}}(x_1) - h_{\{2\}}(x_2) \\
&= a + x_1 + 2x_2 + x_1x_2 - a - x_1 - 2x_2 = x_1x_2.
\end{aligned}
$$

Writing it cleanly, we have:

$$
h_\emptyset = a, \quad h_{\{1\}}(x_1) = x_1, \quad h_{\{2\}}(x_2) = 2x_2, \quad h_{\{1,2\}}(x_1, x_2) = x_1x_2, \tag{8}
$$

It comes as no surprise that in this simple case the fANOVA decomposition does not provide any additional insights as the isolated effects can be directly seen from the function. We show this simple example nevertheless to illustrate at which step which assumption is used. This will make clearer what breaks down when we generalize to dependent variables.

3.1.3 Equality to Hoeffding Decomposition

As we mentioned in section 2 the Hoeffding decomposition laid the groundwork for the fANOVA decomposition. Here, we want to point out that both decompositions yield the same component functions under the assumption of independent and zero-centered inputs. Though we provide no formal proof, we want to illustrate this with our running example (Example 3.1).

Definition 3.2 (Hoeffding decomposition, Il Idrissi et al. (2025))**.** *Let y denote a real-valued function on $\mathbb{R}^N$ with independent inputs $X_1, \ldots, X_N$. The Hoeffding decomposition of y takes the form:*

$$
y(\boldsymbol{X}) = \sum_{A \subseteq D} y_A(\boldsymbol{X}_A), \qquad D := \{1, \ldots, N\}, \tag{9}
$$

where, for each $A \subseteq D$, the component function y_A is defined by:

$$
y_A(\boldsymbol{X}_A) = \sum_{B \subseteq A} (-1)^{|A|-|B|}\, \mathbb{E}[y(\boldsymbol{X}) \mid \boldsymbol{X}_B]\,. \tag{10}
$$

We can apply the Hoeffding decomposition to our running example, without assuming zero mean for now, denote $\mu_1 = \mathbb{E}[X_1]$ and $\mu_2 = \mathbb{E}[X_2]$.

For $A = \emptyset$ there is only one subset $B = \emptyset$. Substituting this into Equation 10 of Definition 3.2 we obtain:

$$h'_{\emptyset} = (-1)^{0-0}\,\mathbb{E}[h(\boldsymbol{X})] = \mathbb{E}[h(\boldsymbol{X})].$$

We compute

$$\mathbb{E}[h(\boldsymbol{X})] = a + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1]\mathbb{E}[X_2] = a + \mu_1 + 2\mu_2 + \mu_1\mu_2.$$

Next, the subsets of $A = \{1\}$ are $B = \emptyset$ and $B = \{1\}$, so

$$h'_{\{1\}}(x_1) = (-1)^{1-0}\,\mathbb{E}[h(\boldsymbol{X})] + (-1)^{1-1}\,\mathbb{E}[h(\boldsymbol{X})|X_1 = x_1] = -\mathbb{E}[h(\boldsymbol{X})] + \mathbb{E}[h(\boldsymbol{X})|X_1 = x_1].$$

Since X_1 is independent of X_2, the conditional expectation is:

$$\mathbb{E}[h(\boldsymbol{X})|X_1 = x_1] = a + x_1 + 2\mu_2 + x_1\mu_2,$$

thus the final expression is given by:

$$h'_{\{1\}}(x_1) = -(a + \mu_1 + 2\mu_2 + \mu_1\mu_2) + (a + x_1 + 2\mu_2 + x_1\mu_2) = (1 + \mu_2)(x_1 - \mu_1).$$

The subsets of $A = \{2\}$ are $B = \emptyset$ and $B = \{2\}$, so

$$h'_{\{2\}}(x_2) = (-1)^{1-0}\,\mathbb{E}[h(\boldsymbol{X})] + (-1)^{1-1}\,\mathbb{E}[h(\boldsymbol{X})|X_2 = x_2] = -\mathbb{E}[h(\boldsymbol{X})] + \mathbb{E}[h(\boldsymbol{X})|X_2 = x_2].$$

Under independence the conditional expectation is:

$$\mathbb{E}[h(\boldsymbol{X})|X_2 = x_2] = a + \mu_1 + 2x_2 + \mu_1 x_2,$$

which yields the expression:

$$h'_{\{2\}}(x_2) = -(a + \mu_1 + 2\mu_2 + \mu_1\mu_2) + (a + \mu_1 + 2x_2 + \mu_1 x_2) = (2 + \mu_1)(x_2 - \mu_2).$$

Finally, the subsets of $A = \{1,2\}$ are $B = \emptyset$, $B = \{1\}$, $B = \{2\}$, $B = \{1,2\}$. We obtain:

$$\begin{aligned} h'_{\{1,2\}}(x_1, x_2) &= (-1)^{2-0}\,\mathbb{E}[h(\boldsymbol{X})] + (-1)^{2-1}\,\mathbb{E}[h(\boldsymbol{X}) \mid X_1 = x_1] \\ &\quad + (-1)^{2-1}\,\mathbb{E}[h(\boldsymbol{X}) \mid X_2 = x_2] + (-1)^{2-2}\,\mathbb{E}[h(\boldsymbol{X}) \mid X_1 = x_1, X_2 = x_2] \\ &= \mathbb{E}[h(\boldsymbol{X})] - \mathbb{E}[h(\boldsymbol{X}) \mid X_1 = x_1] - \mathbb{E}[h(\boldsymbol{X}) \mid X_2 = x_2] \\ &\quad + \mathbb{E}[h(\boldsymbol{X}) \mid X_1 = x_1, X_2 = x_2]. \end{aligned}$$

We already know:

$$\mathbb{E}[h(\boldsymbol{X})|X_1 = x_1, X_2 = x_2] = h(\boldsymbol{X}) = a + x_1 + 2x_2 + x_1x_2.$$

Thus, there interaction term is given by:

$$\begin{aligned} h'_{\{1,2\}}(x_1, x_2) &= (a + \mu_1 + 2\mu_2 + \mu_1\mu_2) - (a + x_1 + 2\mu_2 + \mu_2 x_1) \\ &\quad - (a + \mu_1 + 2x_2 + \mu_1 x_2) + (a + x_1 + 2x_2 + x_1x_2) \\ &= x_1x_2 - \mu_2 x_1 - \mu_1 x_2 + \mu_1\mu_2 \\ &= (x_1 - \mu_1)(x_2 - \mu_2). \end{aligned}$$

Combining these results, the Hoeffding decomposition of $h(x_1, x_2)$ with general means $\mu_1 = \mathbb{E}[X_1]$ and $\mu_2 = \mathbb{E}[X_2]$ is:

$$h'(x_1, x_2) = h'_\emptyset + h'_{\{1\}}(x_1) + h'_{\{2\}}(x_2) + h'_{\{1,2\}}(x_1, x_2),$$

with

$$\begin{aligned} h'_\emptyset &= a + \mu_1 + 2\mu_2 + \mu_1\mu_2, \\ h'_{\{1\}}(x_1) &= (1 + \mu_2)(x_1 - \mu_1), \\ h'_{\{2\}}(x_2) &= (2 + \mu_1)(x_2 - \mu_2), \\ h'_{\{1,2\}}(x_1, x_2) &= (x_1 - \mu_1)(x_2 - \mu_2). \end{aligned}$$

Under the special case of zero-centered input variables, as we assumed in the running example, the decomposition simplifies to:

$$h'_\emptyset = a, \quad h'_{\{1\}}(x_1) = x_1, \quad h'_{\{2\}}(x_2) = 2x_2, \quad h'_{\{1,2\}}(x_1, x_2) = x_1x_2,$$

which coincides with the fANOVA component functions calculated for the polynomial from our running example (Equation 7). The principle of the Hoeffding decomposition is the same as that of the fANOVA decomposition, but the latter is expressed in a recursive form, making explicit that each component accounts for the contributions of lower-order components. In addition, the fANOVA component functions are themselves are zero-centered by construction.

3.1.4 fANOVA via Projection

In the following we revisit the fANOVA decomposition from the view of orthogonal projections. For this section the parallel between the (conditional) expected value and orthogonal projections formulated in Van der Vaart (1998) is crucial. Having this perspective on the fANOVA decomposition helps in bridging different notations of the method (e.g. via expected value or via integral) and also supports in understanding the generalization of fANOVA later in this section. First we define generally what an orthogonal projection is, and then we will use the idea in the context of fANOVA.

Definition 3.3 (Orthogonal Projection, adapted from Nagler (2024a))**.** *Let $\mathcal{G} \subset \mathcal{L}^2$ denote a linear subspace. The projection of y onto $\mathcal{G}$ is defined by the function $\Pi_{\mathcal{G}} y$ which minimizes the distance to y in $\mathcal{L}^2$:*

$$\Pi_{\mathcal{G}} y = \arg\min_{g \in \mathcal{G}} \|y - g\|^2 = \arg\min_{g \in \mathcal{G}} \mathbb{E}[(y(\boldsymbol{X}) - g(\boldsymbol{X}))^2]. \tag{11}$$

When we define the constant component $y_\emptyset$, our goal is to best approximate the original function y by a constant function. In other words, we want to minimize the squared difference between y and a constant function $g_0(x) = a$. The solution is given by the orthogonal projection (see Definition 3.3) of y onto the linear subspace of all constant functions

$$\mathcal{G}_0 = \{g : \Omega \to \mathbb{R} \mid g(x) = a,\ a \in \mathbb{R}\}.$$

In a probabilistic context, we want to minimize the expected squared difference between the random variable $y(\boldsymbol{X})$ and the constant a, which turns out to be equivalent to the expected value of the random variable (Van der Vaart, 1998). Intuitively, in the absence of additional information, the expected value serves as the best approximation of y. More formally, the constant component $y_\emptyset$ is given by:

$$\begin{aligned}\Pi_{\mathcal{G}_0} y &= \arg\min_{g_0 \in \mathcal{G}_0} \|y - g_0\|^2 \\ &= \arg\min_{a \in \mathbb{R}} \mathbb{E}[(y(\boldsymbol{X}) - a)^2] \\ &= \mathbb{E}[y(\boldsymbol{X})] = y_\emptyset.\end{aligned}$$

The main component $y_{\{i\}}(x_i)$ is the projection of y onto the subspace of all functions that only depend on x_i, i.e.,

$$\mathcal{G}_i = \{g : \Omega \to \mathbb{R} \mid g(x) = g_{\{i\}}(x_i)\}.$$

In other words, we minimize the squared difference between y and a function depending only on x_i. The conditional expectation $\mathbb{E}[y(\boldsymbol{X}) \mid X_i = x_i]$ solves this minimization problem (Van der Vaart, 1998), and can at the same time be used to express the fANOVA component functions (Muehlenstaedt et al., 2012). To ensure the interpretation of isolated effects, we subtract lower order terms from the projection. Thus, for the main component $y_{\{i\}}(x_i)$ we have:

$$\begin{aligned}
(\Pi_{\mathcal{G}_i} y)(.) - y_\emptyset &= \arg\min_{g_i \in \mathcal{G}_i} \|y - g_i\|^2 - y_\emptyset \\
&= \arg\min_{g_i \in \mathcal{G}_i} \mathbb{E}\big[(y(\boldsymbol{X}) - g_i(X_i))^2\big] - y_\emptyset \\
&= \mathbb{E}[y(\boldsymbol{X}) \mid X_i = .] - y_\emptyset = y_{\{i\}}(.).
\end{aligned}$$

The second-order interaction component $y_{\{i,j\}}(.,.)$ is the projection of y onto the subspace of all functions that depend on x_i and x_j., i.e.,

$$\mathcal{G}_{i,j} = \{g : \Omega \to \mathbb{R} \mid g(x) = g_{\{i,j\}}(x_i, x_j)\}.$$

Now we are minimizing the squared difference between y and a function depending only on x_i and x_j. Again, we account for effects captured by lower-order components by subtracting the constant and all main components:

$$\begin{aligned}
&(\Pi_{\mathcal{G}_{ij}} y)(.;.) - (y_\emptyset + y_{\{i\}}(.) + y_{\{j\}}(.)) \\
&= \arg\min_{g_{\{i,j\}} \in \mathcal{G}_{ij}} \|y - g_{\{i,j\}}\|^2 - (y_\emptyset + y_{\{i\}}(.) + y_{\{j\}}(.)) \\
&= \arg\min_{g_{\{i,j\}} \in \mathcal{G}_{ij}} \mathbb{E}[(y(\boldsymbol{X}) - g_{\{i,j\}}(.,.))^2] - (y_\emptyset + y_{\{i\}}(.) + y_{\{j\}}(.)) \\
&= \mathbb{E}[y(\boldsymbol{X}) | X_j = ., X_i = .] - (y_\emptyset + y_{\{i\}}(.) + y_{\{j\}}(.)) = y_{\{i,j\}}(.;.)
\end{aligned}$$

In general, for a subset of indices $u \subseteq \{1, \dots, N\}$, we define the subspace

$$\mathcal{G}_u = \{g : \Omega \to \mathbb{R} \mid g(\boldsymbol{x}) = g_u(\boldsymbol{x}_u)\}.$$

Projecting y onto this subspace while subtracting all lower-order components isolates the effect associated exclusively with x_u. This yields the fANOVA component function y_u:

$$\begin{aligned}(\Pi_{\mathcal{G}_u} y)(.) - \sum_{v \subsetneq u} y_v(.) &= \arg\min_{g_u \in \mathcal{G}_u} \|y - g_u\|^2 - \sum_{v \subsetneq u} y_v(.) \\ &= \arg\min_{g_u \in \mathcal{G}_u} \mathbb{E}[(y(\boldsymbol{X}) - g_u(.))^2] - \sum_{v \subsetneq u} y_v(.) \\ &= \mathbb{E}[y(\boldsymbol{X}) \mid X_u = .] - \sum_{v \subsetneq u} y_v(x) = y_u(.).\end{aligned}$$

On this note, we want to highlight that instead of subtracting the lower order components from the projection, it is just as valid to first subtract lower-order components and project y on what is left. We can find both formulations in the literature. For example, Muehlenstaedt et al. (2012) subtracts from the projection and defines:

$$\begin{aligned}y_u(\boldsymbol{x}_u) &:= \mathbb{E}[y(\boldsymbol{X})|\boldsymbol{X}_u = \boldsymbol{x}_u] - \sum_{v \subsetneq u} y_v(\boldsymbol{x}) \\ &= \int_{\mathbb{R}^{N-|u|}} y(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-u}) - \sum_{v \subsetneq u} y_v(\boldsymbol{x}).\end{aligned}$$

Hooker (2004) takes the alternative view and defines the fANOVA component functions via the integral, which can be rewritten as the expected value:

$$\begin{aligned}y_u(\boldsymbol{x}_u) &:= \int_{\mathbb{R}^{N-|u|}} (y(\boldsymbol{x}) - \sum_{v \subsetneq u} y_v(\boldsymbol{x})) d\nu(\boldsymbol{x}_{-u}) \\ &= \mathbb{E}[y(\boldsymbol{X}) - \sum_{v \subsetneq u} y_v(\boldsymbol{x})|\boldsymbol{X}_u = \boldsymbol{x}_u].\end{aligned}$$

The first equivalence in each formulation is simply the definition in each original paper, while the second equivalence holds under the assumption of independent inputs.

3.1.5 Variance Decomposition

Studying the second moments of a function through the lens of the fANOVA decomposition can be useful, especially with regard to the construction of Sobol' indices. We already established that:

$$\mu := \mathbb{E}[y(\boldsymbol{X})] = y_\emptyset.$$

We can also compute the variance of $y(\boldsymbol{X})$ via the fANOVA decomposition. Let the sum over u denote the sum over $\emptyset \neq u \subseteq \{1, \dots, N\}$ and the sum over $u \neq v$ denote the sum over $\emptyset \neq u \subseteq \{1, \dots, N\}$, $\emptyset \neq v \subseteq \{1, \dots, N\}$, $u \neq v$. Following the calculations in

Rahman (2014), the variance of y is then given by:

$$\begin{aligned}\sigma^2 &:= \mathbb{E}\left[(y(\boldsymbol{X}) - \mu)^2\right] \\ &= \mathbb{E}\left[\left(y_\emptyset + \sum_u y_u(\boldsymbol{X}_u) - y_\emptyset\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_u y_u(\boldsymbol{X}_u)\right)^2\right] \\ &= \sum_u \mathbb{E}\left[y_u^2(\boldsymbol{X}_u)\right] + 2\mathbb{E}\left[\sum_{u\neq v} y_u(\boldsymbol{X}_u) y_v(\boldsymbol{X}_v)\right] \\ &= \sum_u \mathbb{E}[y_u^2(\boldsymbol{X}_u)]. \end{aligned} \tag{12}$$

All the cross-terms vanish due to the orthogonality of the fANOVA component functions, i.e. $\mathbb{E}[y_u(\boldsymbol{X}_u)y_v(\boldsymbol{X}_v)] = 0$ for $u \neq v$. This means that the variance of $y(\boldsymbol{X})$ can be decomposed into the sum of the variances of the fANOVA component functions. We verify the variance decomposition for our running example:

$$\operatorname{Var}\big(h(X_1, X_2)\big) = \mathbb{E}\big[h^2_{\{1\}}(X_1)\big] + \mathbb{E}\big[h^2_{\{2\}}(X_2)\big] + \mathbb{E}\big[h^2_{\{1,2\}}(X_1, X_2)\big],$$

where

$$h(x_1, x_2) = a + x_1 + 2x_2 + x_1x_2,$$

and X_1, X_2 are independent with zero mean and unit variance. Starting with the left-hand side and computing the variance of $h(X_1, X_2)$ yields:

$$\begin{aligned}\operatorname{Var}\big(h(X_1, X_2)\big) &= \operatorname{Var}(a + X_1 + 2X_2 + X_1X_2) \\ &= \operatorname{Var}(a) + \operatorname{Var}(X_1) + 4\operatorname{Var}(X_2) + \underbrace{\operatorname{Var}(X_1X_2)}_{(\star)} + 2\operatorname{Cov}(X_1, 2X_2) \\ &= 0 + 1 + 4\cdot 1 + 1\cdot 1 + 2\cdot 0 = 6, \\ (\star)\ \operatorname{Var}(X_1X_2) &= \mathbb{E}[X_1^2X_2^2] - \big(\mathbb{E}[X_1X_2]\big)^2 \\ &= \mathbb{E}[X_1^2]\,\mathbb{E}[X_2^2] - \big(\mathbb{E}[X_1]\,\mathbb{E}[X_2]\big)^2 \\ &= (\operatorname{Var}(X_1) + \mathbb{E}[X_1]^2)\,(\operatorname{Var}(X_2) + \mathbb{E}[X_2]^2) - 0 \\ &= 1\cdot 1 = 1.\end{aligned}$$

For the second line in $(\star)$, we used the fact that independence of any measurable map of X_1 and X_2 follows from independence of X_1 and X_2.

Next, we verify the decomposition from the opposite perspective by starting with the variances of the fANOVA component functions:

$$\mathbb{E}[h^2_{\{1\}}(X_1)] = \mathbb{E}[X_1^2] = 1,$$

$$\mathbb{E}[h^2_{\{2\}}(X_2)] = \mathbb{E}[(2X_2)^2] = 4,$$

$$\mathbb{E}[h^2_{\{1,2\}}(X_1, X_2)] = \mathbb{E}[(X_1X_2)^2] = 1.$$

Combining these expressions, we find:

$$\mathbb{E}[h^2_{\{1\}}(X_1)] + \mathbb{E}[h^2_{\{2\}}(X_2)] + \mathbb{E}[h^2_{\{1,2\}}(X_1, X_2)] = 1 + 4 + 1 = 6 = \mathrm{Var}\left(h(X_1, X_2)\right).$$

3.2 Generalized fANOVA

For the classical fANOVA we make the assumption of independent inputs, which is often violated in practice. In the remainder of this section, we therefore investigate what happens, when we allow for dependency between variables. First, let us recall our running example (see Example 3.1). We modify it slightly by setting $\rho \neq 0$, while keeping everything else the same. When we follow the identical logic as in the classical case we obtain the following terms:

$$
\begin{aligned}
\tilde{h}_\emptyset &= \mathbb{E}[h(X_1, X_2)] \\
&= a + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1 X_2] \\
&= a + \mathbb{E}[X_1 X_2] \\
&= a + \mathrm{Cov}(X_1, X_2) + \mathbb{E}[X_1]\mathbb{E}[X_2] \\
&= a + \rho \\
\tilde{h}_{\{1\}}(x_1) &= \mathbb{E}[h(X_1, X_2)|X_1 = x_1] - \tilde{h}_\emptyset \\
&= \mathbb{E}[a + X_1 + 2X_2 + X_1X_2|X_1 = x_1] - (a + \rho) \\
&= a + x_1 + 2\mathbb{E}[X_2|X_1 = x_1] + x_1\mathbb{E}[X_2|X_1 = x_1] - a - \rho \\
&= x_1 + \rho(2x_1 + x_1^2 - 1) \\
\tilde{h}_{\{2\}}(x_2) &= \mathbb{E}[h(X_1, X_2) \mid X_2 = x_2] - \tilde{h}_\emptyset \\
&= \mathbb{E}[a + X_1 + 2X_2 + X_1X_2 \mid X_2 = x_2] - (a + \rho) \\
&= a + 2x_2 + x_2\mathbb{E}[X_1 \mid X_2 = x_2] - a - \rho \\
&= 2x_2 + \rho(x_2 + x_2^2 - 1) \\
\tilde{h}_{\{1,2\}}(x_1, x_2) &= h(x_1, x_2) - \tilde{h}_\emptyset - \tilde{h}_{\{1\}}(x_1) - \tilde{h}_{\{2\}}(x_2) \\
&= a + x_1 + 2x_2 + x_1x_2 - (a + \rho) \\
&\quad - (x_1 + \rho(2x_1 + x_1^2 - 1)) - (2x_2 + \rho(x_2 + x_2^2 - 1)) \\
&= x_1x_2 - 2\rho x_1 - \rho x_2 - \rho x_1^2 - \rho x_2^2 + \rho
\end{aligned}
$$

The fANOVA component functions are characterized by two central properties: zero-mean and mutual orthogonality, which follow from the strong annihilating conditions. When we check if the components $\tilde{h}_\emptyset, \tilde{h}_{\{1\}}, \tilde{h}_{\{2\}}, \tilde{h}_{\{1,2\}}$ satisfy these properties, we find that all components are zero-centered, but not all are orthogonal to each other. We can, for example, immediately see that checking orthogonality between $\tilde{h}_{\{1\}}, \tilde{h}_{\{1,2\}}$ will yield the expectation over the constant term ρ^2 exactly once, meaning even if all the other expectations cancel out, this constant will remain and the entire expression will be unequal to zero:

$$\begin{aligned}
&\mathbb{E}(\tilde{h}_{\{1\}}(X_1)\tilde{h}_{\{1,2\}}(X_1, X_2)) \\
&= \mathbb{E}[(X_1 + 2\rho X_1 + \rho X_1^2 - \rho) \cdot (X_1 X_2 - 2\rho X_1 - \rho X_2 - \rho X_1^2 - \rho X_2^2 + \rho)] \\
&= \mathbb{E}[X_1^2 X_2] \ldots - \mathbb{E}[\rho^2] \neq 0.
\end{aligned}$$

It turns out that naively computing the "fANOVA decomposition" under dependent inputs, results in components that lack orthogonality, which is a crucial property for interpretability. This shows the need for a more involved approach for a generalization of this method.

3.2.1 Conditions for Generalized fANOVA

Stone (1994) inspired the pioneering work of Hooker (2007) who offered a first solution to the problem of dependent inputs in fANOVA. Work by Chastaing et al. (2012) and Rahman (2014) built on his framework with modifications and extensions. The generalized fANOVA decomposition still follows the overarching form of Definition 3.1. However, we no longer work with a product-type probability measure. Now $f_{\boldsymbol{X}} : \mathbb{R} \to \mathbb{R}_0^+$ denotes an arbitrary probability density function, and we consider the marginal probability measure $f_u(\boldsymbol{x}_u)d\nu(\boldsymbol{x}_u)$ supported on $\mathbb{R}^{|u|}$.

Rather than requiring the strong annihilating conditions (Condition 3.1) for desirable properties of the components, Rahman (2014) proposed to formulate a milder version. The milder version fulfills the same function as the strong version in the classical case but works with the joint density of the variables of interest, instead of the individual marginal probability density functions.

Condition 3.2 (Weak annihilating conditions, (Rahman, 2014))**.** *For the generalized fANOVA decomposition we require, that all the nonconstant fANOVA component functions $y_{u,G}$ integrate to zero w.r.t. the marginal probability density of $\boldsymbol{X}_u$ in each coordinate direction of u, that is,*

$$\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u) d\nu(x_i) = 0 \quad \textit{for} \quad i \in u \neq \emptyset. \tag{13}$$

If components are defined under the weak annihilating conditions, we can ensure that they have zero-mean and exhibit a milder form of orthogonality - hierarchical orthogonality, which means that components of different order are orthogonal to each other while components of the same order are not. Hierarchical orthogonality is the best we can do when independence cannot be assumed.

Proposition 3.3 (Rahman (2014)). *Given the weak annihilating conditions are satisfied, the generalized fANOVA component functions* $y_{u,G}$*, where* $\emptyset \neq u \subseteq \{1, \dots, N\}$*, are centered around zero, i.e.,*

$$\mathbb{E}[y_{u,G}(\boldsymbol{X}_u)] := \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) = 0. \tag{14}$$

Proof. The proof is adjusted from Rahman (2014). For any subset $\emptyset \neq u \subseteq \{1, \dots, N\}$, let $i \in u$. We assume that the weak annihilating conditions are satisfied. Then:

$$\begin{aligned}
\mathbb{E}[y_{u,G}(\boldsymbol{X}_u)] &:= \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^{|u|}} y_{u,G}(\boldsymbol{x}_u) \left(\int_{\mathbb{R}^{N-|u|}} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-u}) \right) d\nu(\boldsymbol{x}_u) \\
&= \int_{\mathbb{R}^{|u|}} y_{u,G}(\boldsymbol{x}_u) f_u(\boldsymbol{x}_u)\, d\nu(\boldsymbol{x}_u) \\
&= \int_{\mathbb{R}^{|u|-1}} \left(\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_u(\boldsymbol{x}_u)\, d\nu(x_i) \right) \prod_{j \in u,\, j \neq i} d\nu(\boldsymbol{x}_j) \\
&= 0,
\end{aligned}$$

where we make use of Fubini's theorem and the last line follows from using the weak annihilating conditions. □

Proposition 3.4 (Rahman (2014)). *Given the weak annihilating conditions are satisfied, two distinct generalized fANOVA component functions* $y_{u,G}$ *and* $y_{v,G}$*, where* $\emptyset \neq u \subseteq \{1, \dots, N\}$*,* $\emptyset \neq v \subseteq \{1, \dots, N\}$*, and* $u \subsetneq v$*, are orthogonal; i.e., they satisfy*

$$\mathbb{E}[y_{u,G}(\boldsymbol{X}_u) y_{v,G}(\boldsymbol{X}_v)] := \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) y_{v,G}(\boldsymbol{x}_v) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) = 0. \tag{15}$$

Proof. The proof is adjusted from Rahman (2014). For any two subsets $\emptyset \neq u \subseteq \{1, \dots, N\}$ and $\emptyset \neq v \subseteq \{1, \dots, N\}$, where $v \subsetneq u$, the subset $u = v \cup (u \setminus v)$. Let

$i \in (u \setminus v) \subseteq u$. Then we obtain:

$$\begin{aligned}
\mathbb{E}[y_{u,G}(\boldsymbol{X}_u)\, y_{v,G}(\boldsymbol{X}_v)] :=& \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) y_{v,G}(\boldsymbol{x}_v) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\
=& \int_{\mathbb{R}^{|u|}} y_{u,G}(\boldsymbol{x}_u) y_{v,G}(\boldsymbol{x}_v) \left(\int_{\mathbb{R}^{N-|u|}} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-u}) \right) d\nu(\boldsymbol{x}_u) \\
=& \int_{\mathbb{R}^{|u|}} y_{u,G}(\boldsymbol{x}_u) y_{v,G}(\boldsymbol{x}_v) f_u(\boldsymbol{x}_u)\, d\nu(\boldsymbol{x}_u) \\
=& \int_{\mathbb{R}^{|v|}} y_{v,G}(\boldsymbol{x}_v) \int_{\mathbb{R}^{|u \setminus v|}} y_{u,G}(\boldsymbol{x}_u) f_u(\boldsymbol{x}_u)\, d\nu(\boldsymbol{x}_{u\setminus v})\, d\nu(\boldsymbol{x}_v) \\
=& \int_{\mathbb{R}^{|v|}} y_{v,G}(\boldsymbol{x}_v) \int_{\mathbb{R}^{|u \setminus v|-1}} \left(\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_u(\boldsymbol{x}_u)\, d\nu(\boldsymbol{x}_i) \right) \\
&\times \prod_{\substack{j \in (u\setminus v) \\ j \neq i}} d\nu(x_j)\, d\nu(\boldsymbol{x}_v) \\
=& \, 0.
\end{aligned}$$

Repeatedly using Fubini's theorem and assuming the weak annihilating conditions are satisfied the equality to zero follows. □

A key contribution from Hooker (2007) and Rahman (2014) is that they construct a generalization of the fANOVA decomposition method as a whole, not only parts, such as generalizing the Sobol' indices. This means it is important that Rahman's generalized statements are coherent with the classical fANOVA decomposition.

Proposition 3.5. *The weak annihilating conditions become the strong annihilating conditions under independence assumption.*

Proof. Assume that the random variables $\{X_j\}_{j \in u}$ are independent. Then we can factorize the marginal density $f_u(\boldsymbol{x}_u)$ as

$$f_u(\boldsymbol{x}_u) = \prod_{j \in u} f_{\{j\}}(x_j).$$

Now we require the weak annihilating conditions for some $i \in u \neq \emptyset$:

$$\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_u(\boldsymbol{x}_u)\, d\nu(x_i) = 0.$$

Since we assume independence, we can substitute the joint marginal density with the product of the marginal densities:

$$\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) \left(\prod_{j \in u} f_{\{j\}}(x_j) \right) d\nu(x_i) = 0.$$

For fixed x_j with $j \neq i$, the term x_i is independent of $f_{\{j\}}(x_j)$, and can therefore be pulled out of the integral:

$$\left(\prod_{j \in u,\, j \neq i} f_{\{j\}}(x_j) \right) \int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_{\{i\}}(x_i)\, d\nu(x_i) = 0.$$

As product of probability density functions the prefactor is strictly positive for all x_j with $j \neq i$. Therefore, the integral must be zero for the equality to hold and we now obtain:

$$\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_{\{i\}}(x_i)\, d\nu(x_i) = 0,$$

which is equivalent to the strong annihilating conditions (Condition 3.1). □

3.2.2 Construction of the Generalized fANOVA Component Functions

Recall the construction of the classical fANOVA component functions (Equation 6). The equation tells us that the nonconstant classical components are defined via the integral of the original function w.r.t. to the product-type probability density function, minus the effects attributed to other components. Ideally, for a well-aligned generalization, we would like the general fANOVA component functions to be built in a similar manner, namely as the integral of y with respect to an appropriately chosen probability density function, minus the effects explained by other components. This is exactly what Rahman (2014) accomplishes. To understand this, we first need to distinguish three cases of integration that will occur in the construction of the generalized components.

Proposition 3.6 (Rahman (2014))**.** *Consider the generalized fANOVA component functions $y_{v,G}$, $\emptyset \neq v \subseteq \{1, \dots, N\}$, of a square-integrable function $y : \mathbb{R}^N \to \mathbb{R}$. When integrated w.r.t. the probability measure $f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u})$, $u \subseteq \{1, \dots, N\}$, one can dis-*

tinguish between three cases:

$$\int_{\mathbb{R}^{N-|u|}} y_{v,G}(\boldsymbol{x}_v) f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u})$$
$$= \begin{cases} \displaystyle\int_{\mathbb{R}^{|v\cap -u|}} y_{v,G}(\boldsymbol{x}_v)\, f_{v\cap -u}(\boldsymbol{x}_{v\cap -u})\, d\nu(\boldsymbol{x}_{v\cap -u}) & \textit{if } v\cap u \neq \emptyset \textit{ and } v \not\subseteq u, \\ y_{v,G}(\boldsymbol{x}_v) & \textit{if } v\cap u \neq \emptyset \textit{ and } v \subseteq u, \\ 0 & \textit{if } v \cap u = \emptyset. \end{cases} \tag{16}$$

Proof. The proof is adjusted from Rahman (2014). Let $u \subseteq \{1,\dots,N\}$ and $\emptyset \neq v \subseteq \{1,\dots,N\}$. We distinguish between three types of relationships between v and u. Before analyzing the first case, note that for any such u and v, it is possible to write

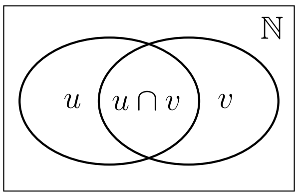
$$(v\cap -u) \subseteq -u \quad \text{and} \quad -u = (-u \setminus (v\cap -u)) \cup (v \cap -u),$$

which will be used in the integral decomposition below.

Case 1: $v\cap u \neq \emptyset$ **and** $v \not\subseteq u$ Next, we use the decomposition of $-u$ stated above to decompose the integration over $\boldsymbol{x}_{-u}$ as:

$$\int_{\mathbb{R}^{N-|u|}} y_{v,G}(\boldsymbol{x}_v)\, f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u})$$
$$= \int_{\mathbb{R}^{|v\cap -u|}} y_{v,G}(\boldsymbol{x}_v) \left(\int_{\mathbb{R}^{N-|u|-|v\cap -u|}} f_{-u}(\boldsymbol{x}_{-u\setminus(v\cap -u)}, \boldsymbol{x}_{v\cap -u})\, d\nu(\boldsymbol{x}_{-u\setminus(v\cap -u)}) \right) d\nu(\boldsymbol{x}_{v\cap -u})$$
$$= \int_{\mathbb{R}^{|v\cap -u|}} y_{v,G}(\boldsymbol{x}_v) f_{v\cap -u}(\boldsymbol{x}_{v\cap -u})\, d\boldsymbol{x}_{v\cap -u},$$

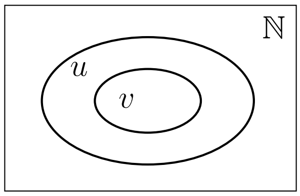
where the inner integral integrates out all variables in $-u \setminus (v\cap -u)$, resulting in the marginal density $f_{v\cap -u}(\boldsymbol{x}_{v\cap -u})$.



Case 2: $v \cap u \neq \emptyset$ **and** $v \subseteq u$**.** Since the sets v and $-u$ are then completely disjoint, $y_{v,G}(\boldsymbol{x}_v)$ is independent of $\boldsymbol{x}_{-u}$ and can be pulled out of the integral:

$$\int_{\mathbb{R}^{N-|u|}} y_{v,G}(\boldsymbol{x}_v) f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}) = y_{v,G}(\boldsymbol{x}_v) \int_{\mathbb{R}^{N-|u|}} f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}) = y_{v,G}(\boldsymbol{x}_v),$$

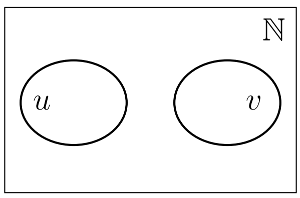
which works because f_{-u} integrates to one.



Case 3: $v \cap u = \emptyset$**.** In this case, we have $v \subseteq -u$, so $v \cap -u = v$. Thus, one can write:

$$\begin{aligned}
\int_{\mathbb{R}^{N-|u|}} y_{v,G}(\boldsymbol{x}_v) f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}) &= \int_{\mathbb{R}^{|v|}} y_{v,G}(\boldsymbol{x}_v) \left(\int_{\mathbb{R}^{N-|u|-|v|}} f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u\setminus v}) \right) d\nu(\boldsymbol{x}_v) \\
&= \int_{\mathbb{R}^{|v|}} y_{v,G}(\boldsymbol{x}_v) f_v(\boldsymbol{x}_v)\, d\nu(\boldsymbol{x}_v) \\
&= \int_{\mathbb{R}^{|v|-1}} \left(\int_{\mathbb{R}} y_{v,G}(\boldsymbol{x}_v) f_v(\boldsymbol{x}_v)\, d\nu(x_i) \right) \prod_{j \in v,\, j \neq i} d\nu(x_j) \\
&= 0,
\end{aligned}$$

while the integral is split in such a way that one recognizes the marginal density f_v, and we employ the zero-mean property.



□

As we will see in the following, we will encounter all of these three integration cases from Proposition 3.6 in the definition of the generalized components by Rahman (2014). In the three integration cases we also already see that the appropriately chosen probability density function is $f_{-u}(\boldsymbol{x}_{-u})$.

Proposition 3.7 (Rahman (2014))**.** *The generalized fANOVA component functions* $y_{u,G}(\boldsymbol{x}_u)$, $u \subseteq \{1, \ldots, N\}$ *of a square-integrable function* $y : \mathbb{R}^N \to \mathbb{R}$ *for a given probability measure*

$f_{\boldsymbol{X}}(\boldsymbol{x})d\nu(\boldsymbol{x})$ *of* $\boldsymbol{X} \in \mathbb{R}^N$ *satisfy*

$$y_{\emptyset,G} = \int_{\mathbb{R}^N} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \tag{17}$$

$$\begin{aligned} y_{u,G}(\boldsymbol{X}_u) = & \int_{\mathbb{R}^{N-|u|}} y(\boldsymbol{X}_u, \boldsymbol{x}_{-u}) f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}) - \sum_{v \subsetneq u} y_{v,G}(\boldsymbol{X}_v) \\ & - \sum_{\substack{\emptyset \neq v \subseteq \{1,\ldots,N\} \\ v\cap u \neq \emptyset,\ v \not\subset u}} \int_{\mathbb{R}^{|v\cap -u|}} y_{v,G}(\boldsymbol{X}_{v\cap u}, \boldsymbol{x}_{v\cap -u}) f_{v\cap -u}(\boldsymbol{x}_{v\cap -u})\, d\nu(\boldsymbol{x}_{v\cap -u}). \end{aligned} \tag{18}$$

Proof. The proof is adjusted from Rahman (2014). We begin by integrating both sides of the generalized fANOVA decomposition

$$y(\boldsymbol{x}) = \sum_{v\subseteq\{1,\ldots,N\}} y_{v,G}(\boldsymbol{x}_v)$$

w.r.t. $f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u})$, replacing $\boldsymbol{X}$ by $\boldsymbol{x}$, and changing the dummy index from u to v. This yields:

$$\int_{\mathbb{R}^{N-|u|}} y(\boldsymbol{x}) f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}) = \sum_{v\subseteq\{1,\ldots,N\}} \int_{\mathbb{R}^{N-|u|}} y_{v,G}(\boldsymbol{x}_v) f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}). \tag{19}$$

Case: $u = \emptyset$. We set $u = \emptyset$, so $-u = \{1, \ldots, N\}$ and $f_{-u}(\boldsymbol{x}_{-u})d\nu(\boldsymbol{x}_{-u}) = f_{\boldsymbol{X}}(\boldsymbol{x})d\nu(\boldsymbol{x})$. The above integral can then be written as:

$$\begin{aligned} \int_{\mathbb{R}^N} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) &= \sum_{v\subseteq\{1,\ldots,N\}} \int_{\mathbb{R}^N} y_{v,G}(\boldsymbol{x}_v) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\ &= y_{\emptyset,G} + \sum_{\emptyset\neq v\subseteq\{1,\ldots,N\}} \int_{\mathbb{R}^N} y_{v,G}(\boldsymbol{x}_v) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\ &= y_{\emptyset,G} + \sum_{\emptyset\neq v\subseteq\{1,\ldots,N\}} \mathbb{E}[y_{v,G}(\boldsymbol{X}_v)] = y_{\emptyset,G}, \end{aligned}$$

where the last sum vanishes given the weak annihilating conditions are satisfied.

Case: $\emptyset \neq u \subseteq \{1, \ldots, N\}$. Returning to the integrated decomposition from Equation 19, we now consider the case where u is non-empty. Making use of the three integration cases from Proposition 3.6 to evaluate each term in the sum according to the relationship between v and u yields four cases:

(A) $v \cap u \neq \emptyset$ and $v \not\subset u$:

This corresponds to case 1 of the three integration cases. The integral becomes:

$$\sum_{\substack{\emptyset \neq v \subseteq \{1,\ldots,N\} \\ v \cap u \neq \emptyset,\ v \not\subseteq u}} \int_{\mathbb{R}^{|v \cap -u|}} y_{v,G}(\boldsymbol{x}_v) f_{v \cap -u}(\boldsymbol{x}_{v \cap -u})\, d\nu(\boldsymbol{x}_{v \cap -u}).$$

(B) $v \subsetneq u$:

This is contained in case 2 of the three integration cases. The integrals reduce to:

$$\sum_{v \subsetneq u} y_{v,G}(\boldsymbol{x}_v).$$

(C) $v = u$:

This is also contained in case 2 of the three integration cases. The integral becomes:

$$y_{u,G}(\boldsymbol{x}_u).$$

(D) $v \cap u = \emptyset$:

This is case 3 of the three integration cases, therefore these terms vanish:

$$\sum_{\substack{v \subseteq \{1,\ldots,N\} \\ v \cap u = \emptyset}} 0 = 0.$$

Combining the above results yields

$$\begin{aligned}\int_{\mathbb{R}^{N-|u|}} y(\boldsymbol{x})\, f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}) = y_{u,G}(\boldsymbol{x}_u) &+ \sum_{v \subsetneq u} y_{v,G}(\boldsymbol{x}_v) \\ &+ \sum_{\substack{\emptyset \neq v \subseteq \{1,\ldots,N\} \\ v \cap u \neq \emptyset,\ v \not\subseteq u}} \int_{\mathbb{R}^{|v \cap -u|}} y_{v,G}(\boldsymbol{x}_v)\, f_{v \cap -u}(\boldsymbol{x}_{v \cap -u})\, d\nu(\boldsymbol{x}_{v \cap -u}).\end{aligned}$$

Rearranging gives the almost final expression for $y_{u,G}(\boldsymbol{x}_u)$:

$$\begin{aligned}y_{u,G}(\boldsymbol{x}_u) = &\int_{\mathbb{R}^{N-|u|}} y(\boldsymbol{x}) f_{-u}(\boldsymbol{x}_{-u})\, d\nu(\boldsymbol{x}_{-u}) \\ &- \sum_{v \subsetneq u} y_{v,G}(\boldsymbol{x}_v) \\ &- \sum_{\substack{\emptyset \neq v \subseteq \{1,\ldots,N\} \\ v \cap u \neq \emptyset,\ v \not\subseteq u}} \int_{\mathbb{R}^{|v \cap -u|}} y_{v,G}(\boldsymbol{x}_v) f_{v \cap -u}(\boldsymbol{x}_{v \cap -u})\, d\nu(\boldsymbol{x}_{v \cap -u}).\end{aligned}$$

As a final step, it remains to write $v = (v \cap u) \cup (v \cap -u)$ to obtain the expression of Proposition 3.7. $\square$

3.2.3 Generalization via Projection

In Hooker (2007) one finds an alternative approach to generalizing the fANOVA decomposition. He extends the framework to dependent input variables by defining the component functions jointly as the solution to a constrained minimization problem:

$$\{y_{u,G}(\boldsymbol{x}_u) \mid u \subseteq d\} = \arg \min_{\{g_u \in \mathcal{L}^2(\mathbb{R}^{|u|})\}} \int_{\mathbb{R}^N} \left(\sum_{u \subseteq d} g_u(\boldsymbol{x}_u) - y(\boldsymbol{x}) \right)^2 f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}), \tag{20}$$

subject to the hierarchical orthogonality conditions:

$$\forall v \subseteq u, \ \forall g_v : \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) g_v(\boldsymbol{x}_v) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) = 0. \tag{21}$$

This formulation is in parallel to projecting y onto a subspace $\mathcal{G}$ under the given constraints (see Definition 3.3).

However, the constraint in Equation 21 is infeasible to enforce in practice. Therefore, Hooker formulates the following proposition, which ensures hierarchical orthogonality of the fANOVA component functions and thus forms the building block of his approach. It can be compared to the weak annihilating conditions (Condition 3.2).

Proposition 3.8 (Hooker (2007))**.** *The hierarchical orthogonality of the fANOVA component functions is ensured if and only if the following integral conditions hold:*

$$\forall u \subseteq N, \ \forall i \in u : \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(x_i) \, d\nu(\boldsymbol{x}_{-u}) = 0. \tag{22}$$

Proof. The proof is adjusted from Hooker (2007). It is organized in two parts. First, Hooker shows that, the hierarchical orthogonality is true, if the integral conditions hold. Second, he shows that hierarchical orthogonality breaks down if the integral conditions are not true. For the first part, assume that the integral conditions from Proposition 3.8 hold. Let $i \in u \setminus v$, then $y_{v,G}(\boldsymbol{x}_v)$ is independent of x_i and $\boldsymbol{x}_{-u}$, which allows us to write:

$$\begin{aligned}
\langle y_{u,G}, y_{v,G} \rangle &:= \int_{\mathbb{R}^N} y_{v,G}(\boldsymbol{x}_v) \, y_{u,G}(\boldsymbol{x}_u) \, f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) \\
&= y_{v,G}(\boldsymbol{x}_v) \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) \, f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) \\
&= 0.
\end{aligned}$$

For the second part, assume that there exists a subset u and an index i for which the integral conditions do not hold, that is,

$$\int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u)\, f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(x_i)\, d\nu(\boldsymbol{x}_{-u}) \neq 0 \quad \text{for some} \quad i, u.$$

Further, assume that hierarchical orthogonality holds for all subsets $v \neq u$ and indices $j \neq i$. Hooker then constructs a fANOVA component function y_v with lower order than y_u, which is not orthogonal to y_u. He sets $v = u \setminus \{i\}$, so $y_{v,G}$ is one order lower than $y_{u,G}$ and defined by:

$$y_{v,G}(\boldsymbol{x}_v) := \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u)\, f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(x_i)\, d\nu(\boldsymbol{x}_{-u}).$$

The constructed $y_{v,G}$ is a valid fANOVA component function, which is unequal to zero by assumption of hierarchical orthogonality being false, while it itself satisfies hierarchical orthogonality by the assumption that

$$\forall j \in v, \quad \int_{\mathbb{R}^N} y_{v,G}(\boldsymbol{x}_v)\, f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(x_j)\, d\nu(\boldsymbol{x}_{-v}) = 0.$$

Lastly, it remains to show, that $y_{v,G}$ is not orthogonal to $y_{u,G}$:

$$\begin{aligned}
\langle y_{u,G}, y_{v,G} \rangle &= \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u)\, y_{v,G}(\boldsymbol{x}_v)\, f_X(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u) \left(\int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u)\, f_X(\boldsymbol{x})\, d\nu(x_i)\, d\nu(\boldsymbol{x}_{-u}) \right) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^N} \left(\int_{\mathbb{R}^N} y_{u,G}(\boldsymbol{x}_u)\, f_X(\boldsymbol{x})\, d\nu(x_i)\, d\nu(\boldsymbol{x}_{-u}) \right)^2 d\nu(\boldsymbol{x}_{u \setminus \{i\}}) \\
&\neq 0.
\end{aligned}$$

□

Hooker approaches his generalization through the lens of projections while Rahman gives a form that tries to imitate the classical fANOVA component functions. A crucial parallel of both versions which distinguishes them from the classical case is that their components are defined in dependence of each other (Proposition 3.7, Equation 20). This makes it in general difficult to compute the generalized fANOVA component functions analytically, even for simple functions.

3.2.4 Generalized Variance Decomposition

Given that the fANOVA decomposition changes under dependent inputs, we briefly make an adjustment to the second-moment statistics of the generalized fANOVA decomposition. The mean of y remains unchanged and is still given by the constant component $y_{\emptyset,G}$, i.e.,

$$\mu_G := \mathbb{E}[y(\boldsymbol{X})] = y_{\emptyset,G}.$$

In contrast, the variance decomposition does not simplify in the same way as in Equation 12, since cross-terms of the same order do not vanish under hierarchical orthogonality. For $\emptyset \neq u \subseteq \{1, \ldots, N\}$, $\emptyset \neq v \subseteq \{1, \ldots, N\}$, $u \neq v$, we restate from above:

$$\begin{aligned}
\sigma^2 &:= \mathbb{E}\left[(y(\boldsymbol{X}) - \mu_G)^2\right] \\
&= \mathbb{E}\left[\left(y_{\emptyset,G} + \sum_u y_{u,G}(\boldsymbol{X}_u) - y_{\emptyset,G}\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_u y_{u,G}(\boldsymbol{X}_u)\right)^2\right] \\
&= \sum_u \mathbb{E}\left[y_{u,G}^2(\boldsymbol{X}_u)\right] + \sum_{u \not\subseteq v,\, v \not\subseteq u} \mathbb{E}\left[y_{u,G}(\boldsymbol{X}_u) y_{v,G}(\boldsymbol{X}_v)\right],
\end{aligned}$$

while the first sum in the final line goes over all nonempty subsets u and the second sum goes over all pairs of subsets (u, v) where neither is a subset of the other one. Conceptually this means that the first term is the sum of the variances of the components, while the second term is the sum of the covariances between components that are not hierarchically orthogonal. The indices under the second component capture precisely the cross-terms that do not vanish under hierarchical orthogonality. As we saw earlier, cross-terms of the same hierarchy also cancel out under the orthogonality assumption of the classical fANOVA.

3.2.5 Example: Dependent Multivariate Normal Inputs

Before ending this section, it remains to answer how the true generalized fANOVA decomposition looks like for our running example. While the interdependence of the generalized components makes it difficult to arrive at an analytical solution, Rahman (2014) provides a way to obtain the closed-form solution for any polynomial of maximum two degree under normally distributed input variables. The approach in Rahman (2014) is based on Fourier–polynomial expansion, which expresses a function as a weighted sum of orthogonal basis functions. This shifts the problem from directly determining the components to identifying the explicit form of the basis functions, through which the generalized fANOVA component functions can be expressed. Under the assumption of normally distributed input variables, Rahman (2014) chooses Hermite polynomials as the basis functions. It remains to determine the weights associated with the basis functions, which can be done via coefficient matching. A polynomial of degree two has the general form:

$$y(x_1, x_2) = a_0 + a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{22}x_2^2 + a_{12}x_1x_2. \tag{23}$$

Any such polynomial may be expressed as a sum of weighted basis functions (see Nagler (2024b)):

$$\begin{aligned} y(x_1, x_2) = c_0 &+ c_{1,1}\,\psi_{1,1}(x_1) + c_{2,1}\,\psi_{2,1}(x_2) \\ &+ c_{1,2}\,\psi_{1,2}(x_1) + c_{2,2}\,\psi_{2,2}(x_2) \\ &+ c_{12,11}\,\psi_{12,11}(x_1, x_2), \end{aligned}$$

where the $\psi_{i,j}$ are the basis functions with corresponding weights $c_0, \dots, c_{12,11} \in \mathbb{R}$. The idea is to carefully construct a set of hierarchically orthogonal basis functions with zero-mean property. Then the expansion in these basis functions is already the fANOVA decomposition of a quadratic polynomial, i.e.,

$$\begin{aligned} y(x_1, x_2) &= a_0 + a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{22}x_2^2 + a_{12}x_1x_2 \\ &= c_0 + c_{1,1}\,\psi_{1,1}(x_1) + c_{2,1}\,\psi_{2,1}(x_2) \\ &\quad + c_{1,2}\,\psi_{1,2}(x_1) + c_{2,2}\,\psi_{2,2}(x_2) + c_{12,11}\,\psi_{12,11}(x_1, x_2) \\ &= \underbrace{c_0}_{y_0} + \underbrace{\big(c_{1,1}\,\psi_{1,1}(x_1) + c_{1,2}\,\psi_{1,2}(x_1)\big)}_{y_1(x_1)} \\ &\quad + \underbrace{\big(c_{2,1}\,\psi_{2,1}(x_2) + c_{2,2}\,\psi_{2,2}(x_2)\big)}_{y_2(x_2)} \\ &\quad + \underbrace{c_{12,11}\,\psi_{12,11}(x_1, x_2)}_{y_{12}(x_1,x_2)}. \end{aligned}$$

We use a slightly simplified version of the basis functions proposed in Rahman (2014) to find an explicit solution for our running example[1]. The basis functions we work with are

[1] We omit the scaling factor, which means the basis functions are not orthonormal anymore but still orthogonal.

given by:

$$\begin{aligned}
\psi_{\emptyset}(x_1, x_2) &= 1, \\
\psi_{1,1}(x_1) &= x_1, \\
\psi_{2,1}(x_2) &= x_2, \\
\psi_{1,2}(x_1) &= x_1^2 - 1, \\
\psi_{2,2}(x_2) &= x_2^2 - 1, \\
\psi_{12,11}(x_1, x_2) &= \frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2},
\end{aligned}$$

where ρ is the correlation coefficient between X_1 and X_2. So this formula will work for dependent as well as independent inputs. What remains it to find the coefficients $c_0, c_{1,1}, \ldots, c_{12,11}$ such that the weighted sum of the basis functions truly recovers the original polynomial (Equation 23). To find the correct weights, we substitute the basis functions and rearrange terms to recognize the groups more easily:

$$\begin{aligned}
y(x_1, x_2) &= c_0 + c_{1,1} x_1 + c_{2,1} x_2 + c_{1,2}(x_1^2 - 1) + c_{2,2}(x_2^2 - 1) \\
&\quad + c_{12,11} \left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2} \right) \\
&= \left(c_0 - c_{1,2} - c_{2,2} + c_{12,11} \tfrac{\rho(\rho^2-1)}{1+\rho^2}\right) + c_{1,1}\, x_1 + c_{2,1}\, x_2 \\
&\quad + \left(c_{1,2} + c_{12,11} \tfrac{\rho}{1+\rho^2}\right) x_1^2 + \left(c_{2,2} + c_{12,11} \tfrac{\rho}{1+\rho^2}\right) x_2^2 - c_{12,11}\, x_1 x_2.
\end{aligned}$$

Now we can use monomial matching to find the coefficients. It is best to start with the interaction term and work backwards from there to the constant term, plugging in the current solutions along the way:

$$\begin{aligned}
-c_{12,11} &= a_{12} & &\Rightarrow & c_{12,11} &= -a_{12} \\
c_{1,2} + c_{12,11} \tfrac{\rho}{1+\rho^2} &= a_{11} & &\Rightarrow & c_{1,2} &= a_{11} + \tfrac{\rho}{1+\rho^2} a_{12} \\
c_{2,2} + c_{12,11} \tfrac{\rho}{1+\rho^2} &= a_{22} & &\Rightarrow & c_{2,2} &= a_{22} + \tfrac{\rho}{1+\rho^2} a_{12} \\
c_{1,1} &= a_1 \\
c_{2,1} &= a_2 \\
c_0 - c_{1,2} - c_{2,2} + c_{12,11} \tfrac{\rho(\rho^2-1)}{1+\rho^2} &= a_0 & &\Rightarrow & c_0 &= a_0 + a_{11} + a_{22} + \rho\, a_{12}
\end{aligned}$$

Hence, the generalized fANOVA decomposition of a two-degree polynomial is given by:

$$
\begin{aligned}
y(x_1, x_2) &= c_0 + c_{1,1}\,\psi_{1,1}(x_1) + c_{2,1}\,\psi_{2,1}(x_2) \\
&\quad + c_{1,2}\,\psi_{1,2}(x_1) + c_{2,2}\,\psi_{2,2}(x_2) + c_{12,11}\,\psi_{12,11}(x_1, x_2) \\
&= \underbrace{\left(a_0 + a_{11} + a_{22} + \rho\, a_{12}\right)}_{c_0} + \underbrace{a_1}_{c_{1,1}} x_1 + \underbrace{a_2}_{c_{2,1}} x_2 \\
&\quad + \underbrace{\left(a_{11} + \frac{\rho}{1+\rho^2} a_{12}\right)}_{c_{1,2}} (x_1^2 - 1) + \underbrace{\left(a_{22} + \frac{\rho}{1+\rho^2} a_{12}\right)}_{c_{2,2}} (x_2^2 - 1) \\
&\quad + \underbrace{(-a_{12})}_{c_{12,11}} \left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2} \right).
\end{aligned}
$$

The individual components are:

$$
\begin{aligned}
y_{\emptyset,G} &= a_0 + a_{11} + a_{22} + \rho\, a_{12}, \\
y_{\{1\},G}(x_1) &= a_1\, x_1 + \left(a_{11} + \frac{\rho}{1+\rho^2} a_{12}\right) \left(x_1^2 - 1\right), \\
y_{\{2\},G}(x_2) &= a_2\, x_2 + \left(a_{22} + \frac{\rho}{1+\rho^2} a_{12}\right) \left(x_2^2 - 1\right), \\
y_{\{1,2\},G}(x_1, x_2) &= -a_{12} \left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2} \right).
\end{aligned}
\tag{24}
$$

This chosen set of basis functions yields valid generalized fANOVA component functions under the assumption of Gaussian inputs. The basis representation is still correct for other distribution assumptions in the sense that it recovers the original function; however, the components would not be hierarchically orthogonal anymore. With this we are able to give the fANOVA component functions for our running example in a generalized form. For $h(x_1, x_2) = x_1 + 2x_2 + x_1 x_2$ we have $a_0 = 0, a_1 = 1, a_2 = 2, a_{11} = 0, a_{22} = 0, a_{12} = 1,$

and therefore obtain with Equation 24:

$$
\begin{aligned}
h_{\emptyset,G} &= \rho, \\
h_{\{1\},G}(x_1) &= x_1 + \frac{\rho}{1+\rho^2}(x_1^2 - 1), \\
h_{\{2\},G}(x_2) &= 2\,x_2 + \frac{\rho}{1+\rho^2}(x_2^2 - 1), \\
h_{\{1,2\},G}(x_1, x_2) &= -\left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2}\right).
\end{aligned}
$$

We refer back to Equation 8 for a comparison with the classical fANOVA components.

4 Visualization and Estimation

In the final section of this thesis, we explore the fANOVA decomposition visually. This provides a better understanding for how the components behave in different scenarios. We first revisit our running example and then explore some other functions.

4.1 Comparison of Decompositions

Recall the polynomial in our running example:

$$h(x_1, x_2) = x_1 + 2x_2 + x_1x_2,$$

with polynomial coefficients: $a_0 = 0$, $a_1 = 1$, $a_2 = 2$, $a_{11} = 0$, $a_{22} = 0$, $a_{12} = 1$. Under independent inputs ($\rho = 0$), the fANOVA component functions are given by:

$$\begin{aligned} h_\emptyset &= 0, \\ h_{\{1\}}(x_1) &= x_1 \\ h_{\{2\}}(x_2) &= 2x_2 \\ h_{\{1,2\}}(x_1, x_2) &= x_1x_2, \end{aligned}$$

visualized in Figure 1. As expected, we observe simple linear functions and a regular symmetric contour plot. Now we assume $\rho = 0.5$. In attempt to compute the fANOVA

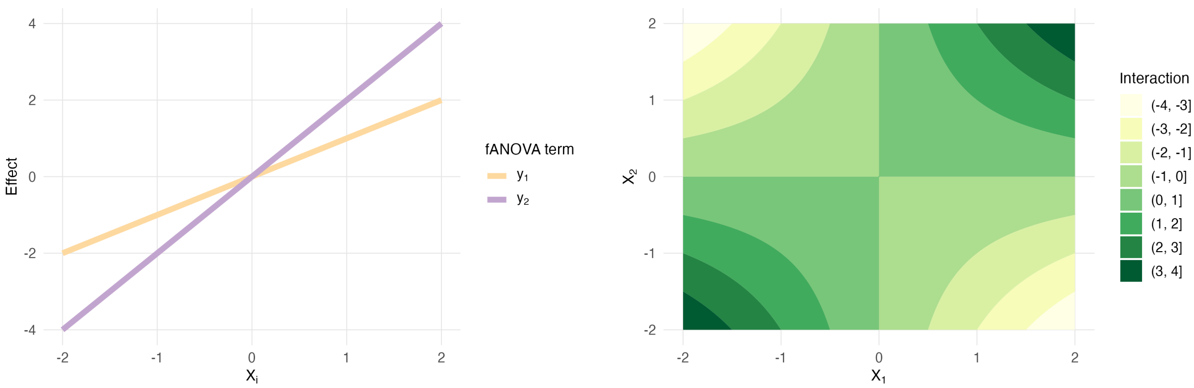


Figure 1: Main fANOVA component functions (left) and interaction component (right) of $h(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with independent inputs.

component functions under dependent inputs, we calculated the following components, which do not satisfy the fANOVA property of orthogonality:

$$
\begin{aligned}
\tilde{h}_0 &= a + 0.5, \\
\tilde{h}_{\{1\}}(x_1) &= 2x_1 + 0.5x_1^2 - 0.5, \\
\tilde{h}_{\{2\}}(x_2) &= 2.5x_2 + 0.5x_2^2 - 0.5, \\
\tilde{h}_{\{1,2\}}(x_1, x_2) &= x_1x_2 - x_1 - 0.5x_2 - 0.5x_1^2 - 0.5x_2^2 + 0.5.
\end{aligned}
$$

Nevertheless, it is interesting to compare their visualization in Figure 2 to the one of the true generalized components in Figure 3. The main effects are parabolic, and the interaction component seems to be non-symmetric.

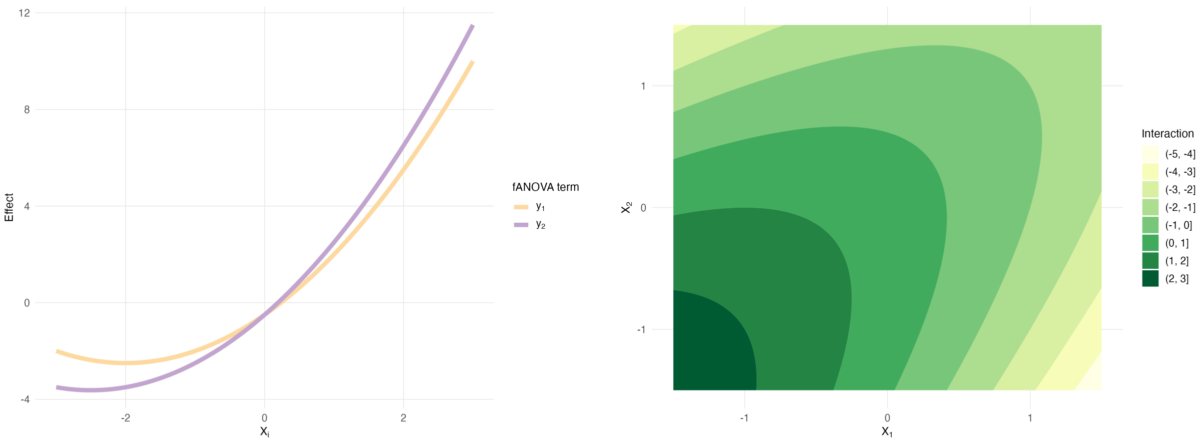


Figure 2: Main effects (left) and interaction effect (right) from a fANOVA-type decomposition of $h(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with dependent inputs, $\rho = 0.5$.

The true fANOVA component functions under $\rho = 0.5$ are given by:

$$
\begin{aligned}
h_{\emptyset,G} &= 0.5, \\
h_{\{1\},G}(x_1) &= x_1 + 0.4\,(x_1^2 - 1) = x_1 + 0.4x_1^2 - 0.4, \\
h_{\{2\},G}(x_2) &= 2x_2 + 0.4\,(x_2^2 - 1) = 2x_2 + 0.4x_2^2 - 0.4, \\
h_{\{1,2\},G}(x_1, x_2) &= -\Big(0.4(x_1^2 + x_2^2) - x_1x_2 - 0.3\Big) \\
&= -0.4x_1^2 - 0.4x_2^2 + x_1x_2 + 0.3.
\end{aligned}
$$

These are visualized in Figure 3. Interestingly, the parabolic form of the main effects is similar between both decompositions, but the interaction effects diverge notably.

Our running example included linear effects of both input variables and an interaction

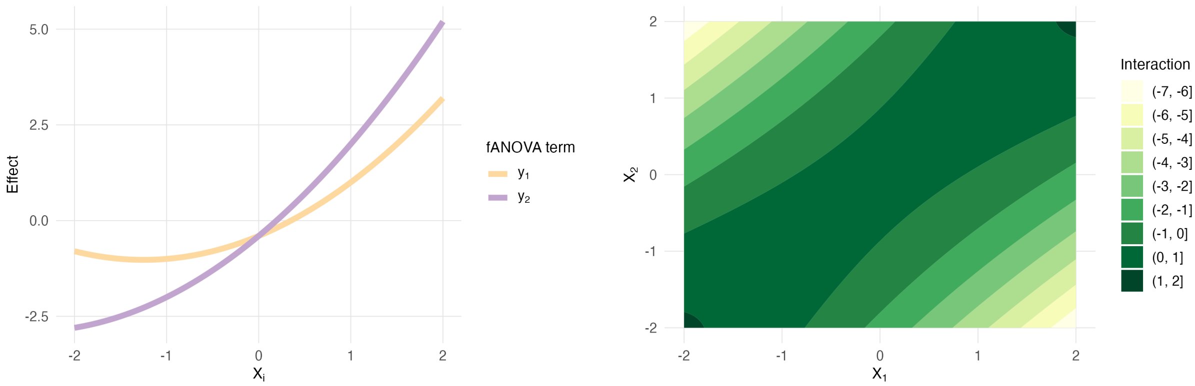


Figure 3: Main fANOVA component functions (left) and interaction component (right) from the generalized fANOVA decomposition of $h(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with dependent inputs, $\rho = 0.5$.

term. For the remainder of this section, we explore other representative scenarios, we can build within the scaffold of a bivariate two-degree polynomial.

4.2 Comparison of Functions

4.2.1 Linear

First, we consider two-degree polynomials of the form:

$$q(x_1, x_2) = a_1x_1 + a_2x_2.$$

We can immediately read off the fANOVA component functions or use the general set of fANOVA components for a two-degree polynomial in Equation 24 which simplify for q to:

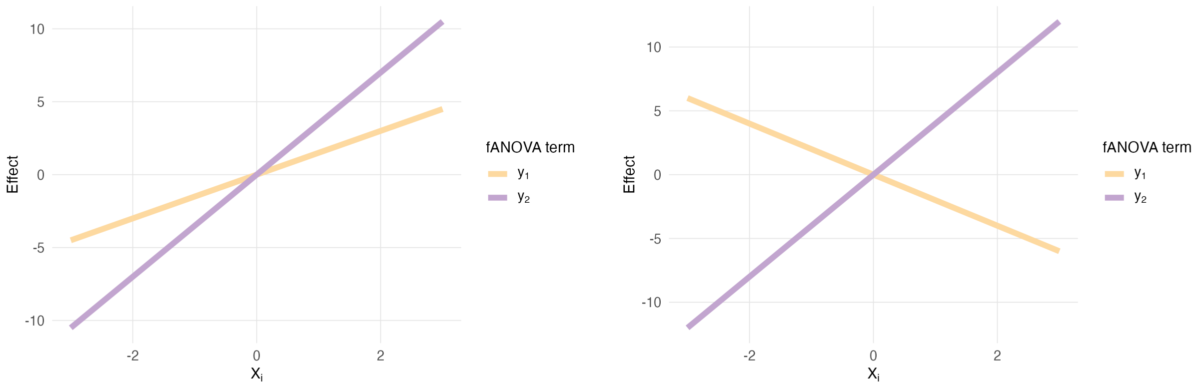
$$q_{\{1\}}(x_1) = a_1x_1,$$
$$q_{\{2\}}(x_2) = a_2x_2.$$

The function q can solely be described by linear main effects (Figure 4). Since no interaction effect is present varying ρ has no impact on the main effects.

4.2.2 Linear and Quadratic

Slightly more complex is a two-degree polynomials, which allows for effects of linear and quadratic nature:

$$p(x_1, x_2) = a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{22}x_2^2.$$



(a) $q(x_1, x_2) = 1.5x_1 + 3.5x_2$ (b) $q(x_1, x_2) = -2x_1 + 4x_2$

Figure 4: Main fANOVA component functions $q_{\{1\}}(x_1) = a_1x_1$ and $q_{\{2\}}(x_2) = a_2x_2$ of the linear function $q(x_1, x_2) = a_1x_1 + a_2x_2$.

The fANOVA component functions for p are given by:

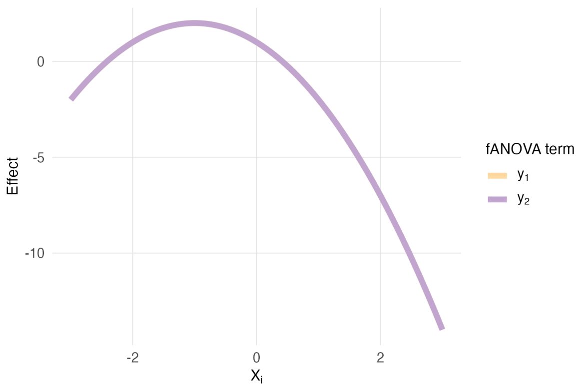
$$
\begin{aligned}
p_\emptyset &= a_{11} + a_{22}, \\
p_{\{1\}}(x_1) &= a_1x_1 + a_{11}(x_1^2 - 1), \\
p_{\{2\}}(x_2) &= a_2x_2 + a_{22}(x_2^2 - 1).
\end{aligned}
$$

We observe parabolic main effects now. In Figure 5, we vary the coefficients a_1, a_2, a_{11}, and a_{22}, while the interaction component is still absent. The coefficients of the quadratic terms determine whether the parabola is facing downwards or upwards; when a_{11} and a_{22} are both negative or both positive the parabola is open downwards or upwards respectively, and when they have opposite signs the parabolas are open in different directions. Alongside the quadratic coefficients, the linear ones a_1 and a_2 influence how stretched or compressed the parabola is.

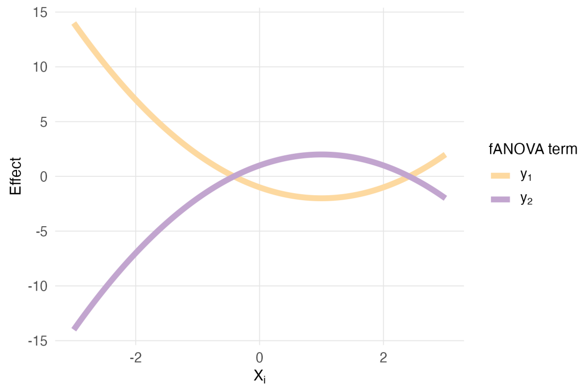
4.2.3 Interaction

Next, we consider a model, which solely consists of an interaction term:

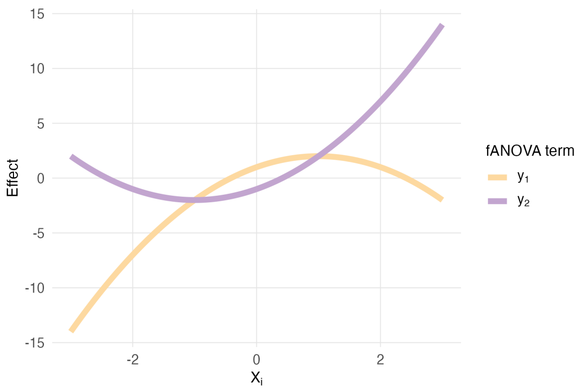
$$w(x_1, x_2) = a_{12}x_1x_2.$$



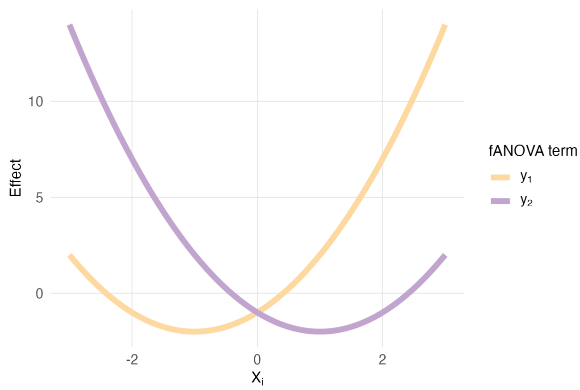
(a) $p(x_1, x_2) = -2x_1 - 2x_2 - x_1^2 - x_2^2$



(b) $p(x_1, x_2) = -2x_1 + 2x_2 + x_1^2 - x_2^2$



(c) $p(x_1, x_2) = 2x_1 + 2x_2 - x_1^2 + x_2^2$



(d) $p(x_1, x_2) = 2x_1 - 2x_2 + x_1^2 + x_2^2$

Figure 5: Main fANOVA component functions $p_{\{1\}}(x_1) = a_1x_1 + a_{11}(x_1^2 - 1)$ and $p_{\{2\}}(x_2) = a_2x_2 + a_{22}(x_2^2 - 1)$ of the polynomial $p(x_1, x_2) = a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{22}x_2^2$.

The fANOVA component functions for w are given by:

$$\begin{aligned}
w_{\emptyset,G} &= a_{12}\rho, \\
w_{\{1\},G}(x_1) &= a_{12}\frac{\rho}{1+\rho}(x_1^2-1), \\
w_{\{2\},G}(x_2) &= a_{12}\frac{\rho}{1+\rho}(x_2^2-1), \\
w_{\{1,2\},G}(x_1,x_2) &= -a_{12}\left(\frac{\rho(x_1^2+x_2^2)}{1+\rho^2} - x_1x_2 + \frac{\rho(\rho^2-1)}{1+\rho^2}\right).
\end{aligned}$$

The main components $w_{\{1\},G}$ and $w_{\{2\},G}$, as well as the interaction component $w_{\{1,2\},G}$, are influenced by ρ and a_{12}. In our example we keep $a_{12} = 2$ fixed and show the interaction effect as a contour plot for varying ρ with the corresponding main effects next to it in Figure 6. The main effects have the same form for every case of ρ and a_{12} and thus overlap.
This example is simple yet interesting because it shows that in the case where the true function consists solely of an interaction term, fANOVA still attributes something to the isolated effect of each variable. Only when the variables are uncorrelated, all the effect is attributed to the interaction term. This functionality hints to why Lengerich et al. (2020) build an algorithm around fANOVA to purify interaction effects[2].

4.2.4 Full

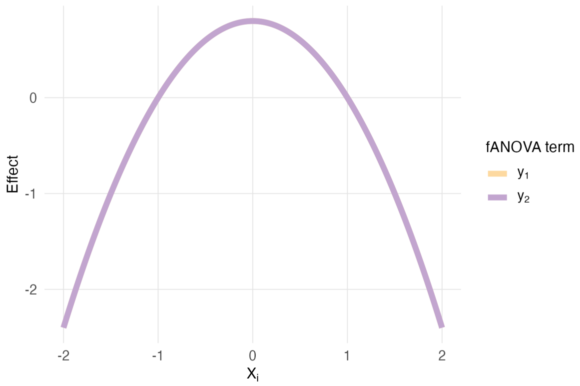
Finally, we consider a full example, including all main and interaction effects:

$$z(x_1,x_2) = a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{22}x_2^2 + a_{12}x_1x_2.$$

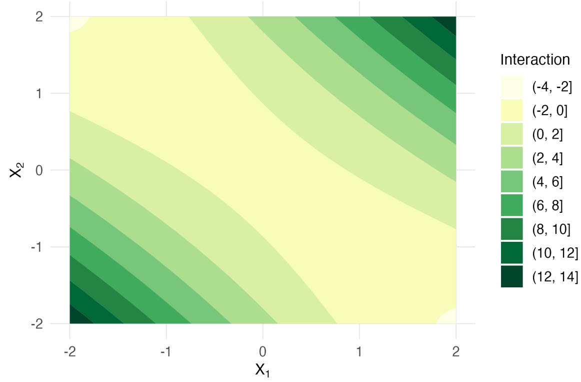
Now the fANOVA component functions are given by Equation 24, where $a_0 = 0$. We can vary the coefficients as well as ρ.

When the true function has no interaction term, as in our first two scenarios, varying ρ is uninteresting because there is no way it could influence the form of the main effects. In this full scenario, however, there is an interaction term present, and therefore it is most interesting to compare pairs of coefficient sets under $\rho = 0$ versus $\rho \neq 0$. With this we want to essentially ask how effects are distorted by performing the classical fANOVA decomposition when a true interaction term is present and variables exhibit dependency. In Figure 7 we make this comparison for a weakly positive linear correlation between

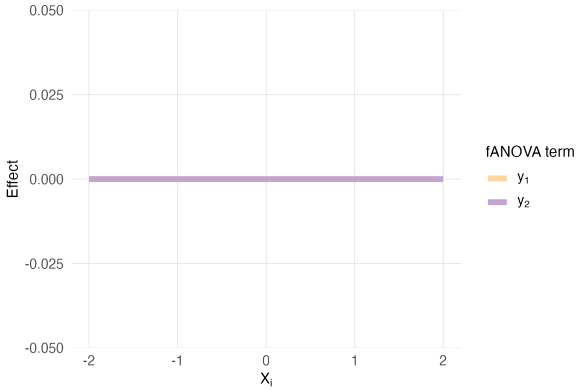
[2]Because they see a pure interaction as an effect which cannot be attributed to lower order terms; this means when identifying interactions we want to attribute all we can to lower order terms and what is left is the true interaction effect.



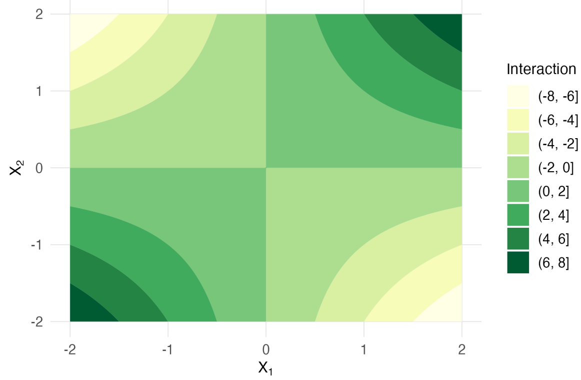
(a) Main effects for $\rho = -0.5$



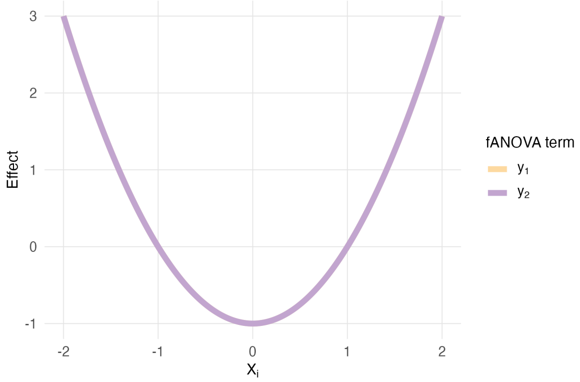
(b) Interaction effect for $\rho = -0.5$



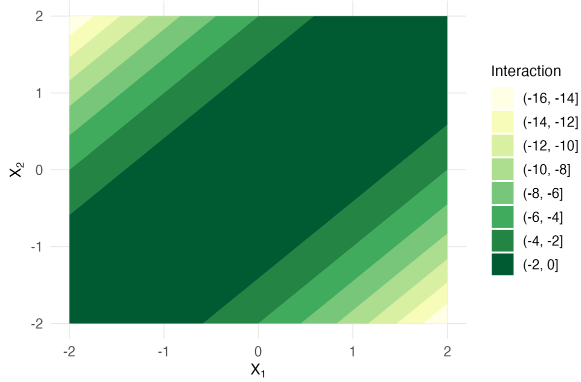
(c) Main effects for $\rho = 0$



(d) Interaction effect for $\rho = 0$



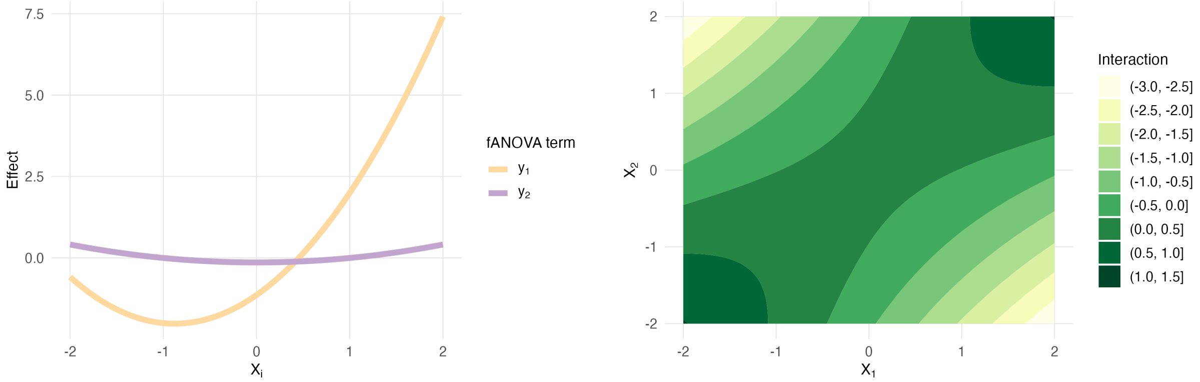
(e) Main effects for $\rho = 1$



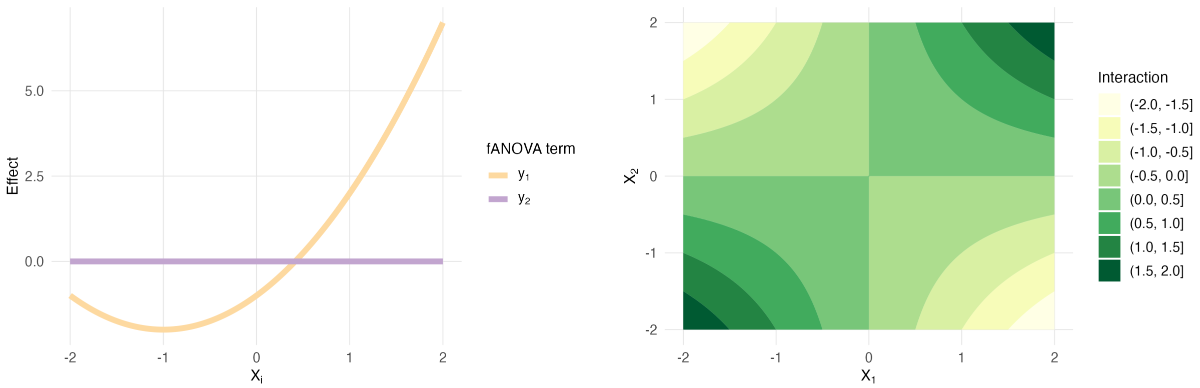
(f) Interaction effect for $\rho = 1$

Figure 6: Main fANOVA component functions (left) and interaction component (right) of the function $w(x_1, x_2) = 2x_1x_2$ for varying ρ.

variables and in Figure 8 we show the same for a strongly negative linear correlation between variables. Similar to the visualization of our running example in Figure 3, we see that main effects are distorted slightly, while interaction effects look substantially different under dependent inputs.



(a) Main effect for $\rho = 0.3$. (b) Interaction effect for $\rho = 0.3$.



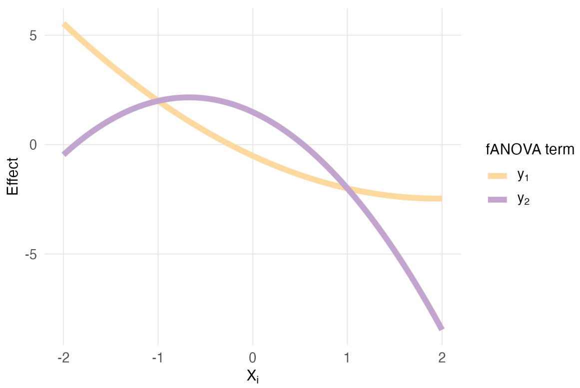
(c) Main effect for $\rho = 0$. (d) Interaction effect for $\rho = 0$.

Figure 7: Main fANOVA component functions (left) and interaction component (right) of the polynomial $z(x_1, x_2) = 2x_1 + x_1^2 + 0.5x_1x_2$ for varying ρ. The coefficient sets are identical while the correlation structure varies.

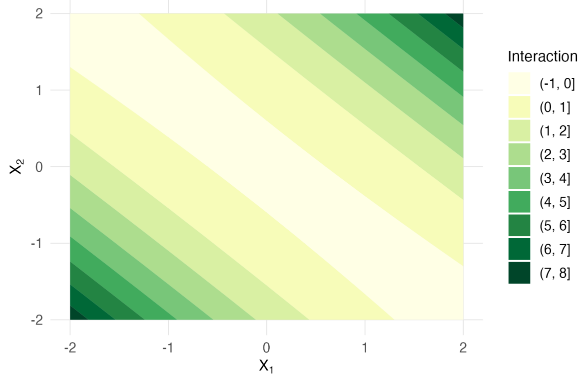
4.3 Estimation of fANOVA Component Functions

The discussion so far has been mostly theoretical, and the examples used were deliberately simple to illustrate the key ideas. In practical applications, however, the true function is typically unknown and more complex. This makes the development of suitable estimation procedures an important step towards establishing the method as a practical interpretability tool.

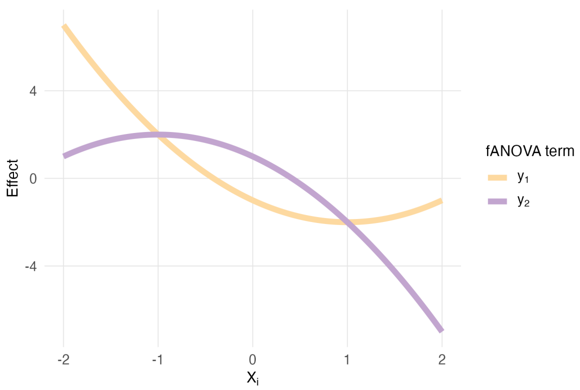
We already encountered one estimation scheme proposed by Rahman (2014) when computing the generalized fANOVA component functions for our running example; we



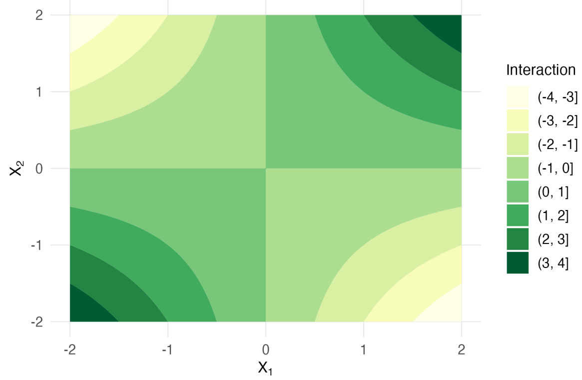
(a) Main effect for $\rho = -0.8$.



(b) Interaction effect for $\rho = -0.8$.



(c) Main effect for $\rho = 0$.



(d) Interaction effect for $\rho = 0$.

Figure 8: Main fANOVA component functions (left) and interaction component (right) of the polynomial $z(x_1, x_2) = -2x_1 - 2x_2 + x_1^2 - x_2^2 + x_1x_2$ for varying ρ. The coefficient sets are identical, while the correlation structure varies.

refer to section 3 for the conceptual idea behind it.

In Hooker (2004), an estimation framework based on partial dependence was proposed, which makes use of the formulation of fANOVA via projections. To obtain the component estimate for y_u, Hooker proposes to estimate the projections of y onto the subspace of variables spanned by u empirically. One does so by first estimating the conditional expected value of the variables in u. This is a simple Monte Carlo estimation, which results in the partial dependence function (PD Function) for the variables in u (Hooker, 2004). The PD Function is then used to estimate the empirical projection of interest. He states that his method works well for functions that truly have nearly additive structure, and purely additive functions are exactly recoverable with this approach. However, the approach suffers from extrapolation issues or artefacts when the true function involves interactions and inputs are dependent.

Therefore, in Hooker (2007), a new estimation scheme was proposed for his version of the generalized fANOVA decomposition (see section 3). Hooker rewrites his proposed system of equations as a restricted weighted least squares problem and solves it via a Lagrange multiplier for the exact solution of the simultaneously defined generalized components. The function is evaluated at a grid of points to reduce computational costs. Because of its parallel to weighted least squares, it is possible to compute a weighted standard ANOVA using existing software. However, this approach makes it difficult to incorporate system constraints, and the resulting components may fail to be hierarchically orthogonal.

5 Conclusion

The fANOVA decomposition is a foundational method that has been studied from many perspectives, with recent interest driven by its potential in model interpretability. Its key strength lies in producing components that represent the unique contribution of each variable without mixing effects.

Despite being established in the literature, we found a lack of unified formalization and notational clarity, which we aimed to address in this work.

We began by introducing the fANOVA decomposition under the strong annihilating conditions, resulting in zero-mean and mutually orthogonal component functions. We filled small gaps in mathematical proofs and illustrated the parallel to the Hoeffding decomposition under independent and zero-centered inputs. Further, we established a direct connection to conditional expectations and orthogonal projections, which we argue is key to unifying different existing notations and formal approaches. Alongside the functional decomposition, we presented the variance decomposition, which underpins Sobol' indices.

Next, we extended the decomposition to dependent input variables. Here, multiple approaches exist; we focused on the frameworks of Hooker (2007) and Rahman (2014). Using an illustrative example, we demonstrated that obtaining interpretable components under dependence requires the careful construction of specific constraints. We also noted that closed-form solutions for the generalized components are difficult to obtain in practice and remain an open problem. Nevertheless, the associated variance decomposition still holds, allowing for the construction of generalized Sobol' indices.

To complement the theoretical work, we visualized fANOVA component functions. These visualizations illustrated how fANOVA separates effects, but were limited to simple functions and Gaussian inputs.

This work did not present a full treatment of Sobol' indices, and only touched on variance decomposition as their foundation. Our empirical demonstrations were restricted to toy examples; applying fANOVA in practice will require estimation methods for trained models on real data. We also briefly mentioned the ability of fANOVA to purify interaction effects (Lengerich et al., 2020) and its relation to other interpretability techniques (e.g., PD, Shapley (Fumagalli et al., 2025)), but did not explore these connections in depth.

Future work could extend Rahman's Fourier-polynomial expansion approach to more complex polynomials and non-Gaussian distributions. Additionally, it would be interesting to investigate mathematical connections between fANOVA and other interpretability frameworks. Lastly, a lack of standard software for performing fANOVA remains a barrier to its widespread adoption, so developing robust, open-source implementations would be an important step toward enabling its use in practical model interpretability.

References

Borgonovo, E., Li, G., Barr, J., Plischke, E. and Rabitz, H. (2022). Global Sensitivity Analysis with Mixtures: A Generalized Functional ANOVA Approach, *Risk Analysis* **42**(2): 304–333.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/risa.13763*

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.

Choi, Y., Park, S., Park, C., Kim, D. and Kim, Y. (2025). Meta-anova: screening interactions for interpretable machine learning, *Journal of Korean Statistical Society* .
URL: *https://link.springer.com/10.1007/s42952-024-00302-2*

Fumagalli, F., Muschalik, M., Hüllermeier, E., Hammer, B. and Herbinger, J. (2025). Unifying Feature-Based Explanations with Functional ANOVA and Cooperative Game Theory. arXiv:2412.17152 [cs].
URL: *http://arxiv.org/abs/2412.17152*

Gu, C. (2013). *Smoothing Spline ANOVA Models*, Vol. 297 of *Springer Series in Statistics*, Springer New York, New York, NY.
URL: *https://link.springer.com/10.1007/978-1-4614-5369-7*

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution, *The Annals of Mathematical Statistics* **19**(3): 293–325. Publisher: Institute of Mathematical Statistics.
URL: *https://www.jstor.org/stable/2235637*

Hooker, G. (2004). Discovering additive structure in black box functions, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, Seattle WA USA, pp. 575–580.
URL: *https://dl.acm.org/doi/10.1145/1014052.1014122*

Hooker, G. (2007). Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables, *Journal of Computational and Graphical Statistics* **16**(3): 709–732.
URL: *http://www.tandfonline.com/doi/abs/10.1198/106186007X237892*

Hu, L., Nair, V. N., Sudjianto, A., Zhang, A., Chen, J. and Yang, Z. (2025). Interpretable Machine Learning Based on Functional ANOVA Framework: Algorithms and Comparisons, *Applied Stochastic Models in Business and Industry* **41**(1): e2916.
URL: *https://onlinelibrary.wiley.com/doi/10.1002/asmb.2916*

Il Idrissi, M., 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.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0047259X25000399*

König, G., Günther, E. and Luxburg, U. v. (2024). Disentangling Interactions and Dependencies in Feature Attribution. arXiv:2410.23772.
URL: *http://arxiv.org/abs/2410.23772*

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, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 2402–2412. ISSN: 2640-3498.
URL: *https://proceedings.mlr.press/v108/lengerich20a.html*

Liu, R. and Owen, A. B. (2006). Estimating Mean Dimensionality of Analysis of Variance Decompositions, *Journal of the American Statistical Association* **101**(474): 712–721.
URL: *https://www.tandfonline.com/doi/full/10.1198/016214505000001410*

Molnar, C. (2025). *Interpretable Machine Learning*, 3 edn.
URL: *https://christophm.github.io/interpretable-ml-book*

Muehlenstaedt, T., Roustant, O., Carraro, L. and Kuhnt, S. (2012). Data-driven Kriging models based on FANOVA-decomposition, *Statistics and Computing* **22**(3): 723–738.
URL: *http://link.springer.com/10.1007/s11222-011-9259-7*

Nagler, T. (2024a). Mathematical statistics, Lecture Script. `https://tnagler.github.io/mathstat-lmu-2024.pdf`.
URL: *https://tnagler.github.io/mathstat-lmu-2024.pdf*

Nagler, T. (2024b). Methoden der Linearen Algebra in der Statistik – Vorlesungsskript, `https://tnagler.github.io/linalg-2024.pdf`. Version Sommersemester 2024.

Owen, A. B. (2013). Variance Components and Generalized Sobol' Indices, *SIAM/ASA Journal on Uncertainty Quantification* **1**(1): 19–41. tex.eprint: https://doi.org/10.1137/120876782.
URL: *https://doi.org/10.1137/120876782*

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

Rahman, S. (2014). A Generalized ANOVA Dimensional Decomposition for Dependent Probability Measures, *SIAM/ASA Journal on Uncertainty Quantification* **2**(1): 670–697.
URL: *https://doi.org/10.1137/120904378*

Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Mathematics and Computers in Simulation* **55**(1-3): 271–280.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0378475400002706*

Sobol, I. M. (1993). Sensitivity Estimates for Nonlinear Mathematical Models, *Mathematical Modelling and Computational Experiments* **1**: 407–414.

Stone, C. J. (1994). The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation, *The Annals of Statistics* **22**(1): 118–171. Publisher: Institute of Mathematical Statistics.
URL: *https://www.jstor.org/stable/2242446*

Stone, C. J., Hansen, M. H., Kooperberg, C. and Truong, Y. K. (1997). Polynomial Splines and their Tensor Products in Extended Linear Modeling, *The Annals of Statistics* **25**(4): 1371–1425. Publisher: Institute of Mathematical Statistics.
URL: *https://www.jstor.org/stable/2959054*

Takemura, A. (1983). Tensor Analysis of ANOVA Decomposition, *Journal of the American Statistical Association* **78**(384): 894–900. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
URL: *https://www.jstor.org/stable/2288201*

Van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

A Appendix

Proof of classical fANOVA decomposition

Here we show the adjusted proof of Theorem 1 in Sobol (1993).

Theorem A.1. *Any function y, which is integrable over the unit hypercube $[0,1]^k$, has a unique fANOVA expansion of the form:*

$$y(\boldsymbol{X}) = \sum_{u \subseteq \{1,\ldots,N\}} y_u(\boldsymbol{X}_u),$$

subject to the constraint that Proposition 3.1 is satisfied.

Sobol' proofs existence and uniqueness of the fANOVA decomposition by showing how the summands of the desired decomposition look and constructing them in such a way that they have the zero-mean property.

Proof. Assume that $\boldsymbol{X}$ is an N-dimensional vector of independent random variables and that the still unspecified fANOVA components have zero-mean. He defines the integral w.r.t. all variables except for the ones with indices in v:

$$g_v(\boldsymbol{x}_v) = \int_{[0,1]^{N-|v|}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-v}).$$

He then builds the fANOVA terms subsequently and shows that they indeed satisfy the desired properties.
The very first term in the decomposition is the integral of y with respect to all variables:

$$y_\emptyset = \int_{\mathbb{R}^N} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}).$$

This integral exists because $y \in \mathcal{L}^2(\mathbb{R}^N, f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}))$, and the product measure is finite on the domain.
Next, Sobol' derives the one-dimensional fANOVA terms. For this, he takes the integral

of Definition 3.1 w.r.t. all variables except for the one with index i, so $v_1 = \{i\}$:

$$\begin{aligned}\int_{\mathbb{R}^{N-1}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) &= \int_{\mathbb{R}^{N-1}} \sum_{u\subseteq\{1,\dots,N\}} y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) \\ &= \sum_{u\subseteq\{1,\dots,N\}} \int_{\mathbb{R}^{N-1}} y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) \\ &= \sum_{u\subseteq\{1,\dots,N\}} \int_{\mathbb{R}^{N-1}} y_u(\boldsymbol{x}_u) \left(\prod_{j=1}^{N} f_{\{j\}}(x_j)\right) d\nu(\boldsymbol{x}_{-v_1}).\end{aligned}$$

For every summand $y_u(\boldsymbol{x}_u)$ with $u \not\ni i$, the integrand does not depend on x_i, and thus vanished due to the zero-mean constraint. Similarly, for any term $y_u(\boldsymbol{x}_u)$ with $i \in u$ and $|u| > 1$, the integration will include at least one other variable in u, again causing the integral to vanish. In the end, only the constant term $y_\emptyset$ and the one-dimensional term $y_{\{i\}}(x_i)$ remain, which depend only on x_i and are not integrated. Therefore, we can derive the simplified expression:

$$\int_{\mathbb{R}^{N-1}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) = y_\emptyset + y_{\{i\}}(x_i).$$

This equation allows to define the one-dimensional term $y_{\{i\}}$ explicitly as:

$$y_{\{i\}}(x_i) = \int_{\mathbb{R}^{N-1}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) - y_\emptyset.$$

Next, he considers $v_2 = \{i, j\}$. The ANOVA decomposition is integrated over all variables except x_i and x_j:

$$\begin{aligned}\int_{\mathbb{R}^{N-2}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-\{i,j\}}) &= \int_{\mathbb{R}^{N-2}} \sum_{u\subseteq\{1,\dots,N\}} y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-\{i,j\}}) \\ &= \sum_{u\subseteq\{1,\dots,N\}} \int_{\mathbb{R}^{N-2}} y_u(\boldsymbol{x}_u) \left(\prod_{j=1}^{N} f_{X_j}(x_j)\right) d\nu(\boldsymbol{x}_{-\{i,j\}}) \\ &= y_\emptyset + y_{\{i\}}(x_i) + y_{\{j\}}(x_j) + y_{\{i,j\}}(x_i, x_j).\end{aligned}$$

Hence, the two-dimensional components are given by:

$$y_{\{i,j\}}(x_i, x_j) = \int_{\mathbb{R}^{N-2}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-\{i,j\}}) - y_\emptyset - y_{\{i\}}(x_i) - y_{\{j\}}(x_j).$$

One can continue this process for all combinations of indices $v \subseteq \{1, \dots, N\}$ to derive

the corresponding fANOVA terms $y_v(\boldsymbol{x}_v)$.
Now let $v \subseteq \{1, \ldots, N\}$. The general expression for the component $y_v(\boldsymbol{x}_v)$ is given by:

$$y_v(\boldsymbol{x}_v) = \int_{\mathbb{R}^{N-|v|}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v}) - \sum_{u \subsetneq v} y_u(\boldsymbol{x}_u).$$

The last term is the decomposition integrated with respect to no variables, i.e., the function itself:

$$y_{\{1,\ldots,N\}}(\boldsymbol{x}) = y(\boldsymbol{x}) - \sum_{u \subsetneq \{1,\ldots,N\}} y_u(\boldsymbol{x}_u).$$

Finally, it remains to verify that the constructed component functions satisfy the zero-mean constraint. Let $v \subseteq \{1, \ldots, N\}$, and let $i \in v$. Then:

$$\begin{aligned}
&\int y_v(\boldsymbol{x}_v) f_{X_i}(x_i)\, d\nu(x_i) \\
&= \int \left(\int y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-v}) - \sum_{u \subsetneq v} y_u(\boldsymbol{x}_u) \right) f_{\{i\}}(x_i)\, d\nu(x_i) \\
&= \int \left(\int y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-v}) \right) f_{\{i\}}(x_i)\, d\nu(x_i) - \sum_{u \subsetneq v} \int y_u(\boldsymbol{x}_u) f_{\{i\}}(x_i)\, d\nu(x_i) \\
&= \int y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-v}) d\nu(x_i) - \sum_{u \subsetneq v} \int y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u)\, d\nu(\boldsymbol{x}_u)
\end{aligned}$$

The first term integrates out all of $\boldsymbol{x}_v$, leaving $y_\emptyset$. Each term in the sum vanishes by the zero-mean property of lower-order components:

$$\int y_v(\boldsymbol{x}_v) f_{\{i\}}(x_i)\, d\nu(x_i) = y_\emptyset - y_\emptyset = 0.$$

Thus, every component $y_v(\boldsymbol{x}_v)$ satisfies:

$$\int y_v(\boldsymbol{x}_v) f_{\{i\}}(x_i)\, d\nu(x_i) = 0, \quad \text{for all } i \in v.$$

□

B Electronic appendix

Data, code and figures are provided in electronic form at:
`https://github.com/juliet-fleischer/fANOVA_decomposition`

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

AI Usage Disclosure

In the preparation of this thesis, AI tools were employed in the following ways:

- Literature Research: ChatGPT-4o, Elicit, and Connected Papers were used to assist in identifying, exploring, and understanding relevant scientific literature.
- Conceptual Support and Mathematical Proofs: ChatGPT-4o was used to clarify complex concepts and to aid in the development and verification of mathematical proofs.
- LaTeX Support: GitHub Copilot and ChatGPT-4o were used for generating and formatting LaTeX formulas.
- Code Assistance: GitHub Copilot and OpenAI Codex, integrated into RStudio and Visual Studio Code, were used to assist with writing and optimizing code.
- Language and Grammar: ChatGPT-4o was used to improve language clarity and grammar, with explicit instructions to preserve the original meaning and content of the text.

Munich, 31 July 2025

Juliet Fleischer