Bachelor's Thesis

---

fANOVA for Interpretable Machine Learning

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, Month Day[th], Year

Abstract

This article studies the functional ANOVA decomposition (fANOVA) in the context of model interpretability. We start with the historical background and providing context to how the method has evolved over time and is used in different domains. This is followed by the formal definition of the method, where we distinguish between the classical fANOVA, which assumes independent inputs, and the generalized fANOVA, which allows for dependency between variables. For both cases, we unite different notations and approaches to the formalization of fANOVA. When generalizing fANOVA to dependent inputs, we encounter the Hoeffding decomposition, which is closely related to fANOVA. Illustrated by examples we clearly distinguish between both methods and experiment with visualizing the decomposition of different functions. We conclude with some words on current estimation approaches and suggestions for future work.

Contents

1 Introduction

Interpretability and sensitivity analysis are increasingly important in machine learning (ML) and in the study of complex systems. One of the foundational mathematical tools supporting these goals is the functional ANOVA decomposition (fANOVA).

At its core, the fANOVA decomposition provides a method which allows decomposing integrable functions into a sum of orthogonal components. It is a foundational method, useful in interpretability of black box models (Hooker (2004), Molnar (2025)), 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 different 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, 1993), functional ANOVA decomposition (Hooker, 2004), ANOVA dimensional decomposition (Rahman, 2014) – in this thesis we will refer to it as the fANOVA decomposition. The diversity does not stop at naming. Different authors formalize the decomposition using different 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)). It is being used to build inherently interpretable models, 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 historical context and related work. This

is followed by the central part in which we give the formal definition of the classical and generalized fANOVA decomposition. Next we illustrate characteristics of the method based on analytical examples, before briefly outlining current estimation schemes and concluding with a discussion and possible future research directions.

2 Prerequisites

Basic Setup

The formal setup is based on Chastaing et al. (2012) and Rahman (2014). 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 a $\mathbb{R}^N$-valued random vector.

We assume that the probability distribution of $\boldsymbol{X}$ is continuous and completely defined by the joint probability density function (pdf) $f_{\boldsymbol{X}} : \mathbb{R}^N \to \mathbb{R}_0^+$. $f_{\boldsymbol{X}}$ is the pdf w.r.t. measure ν.

Let u denote a subset of indices $\{1, \ldots, N\}$, and $-u := \{1, \ldots, N\} \backslash u$ its compliment. We denote strict inclusion by $\subsetneq$ and $\subset$ allows for equality. $\boldsymbol{X_u} = (X_1, \ldots, X_{|u|}), u \neq \emptyset, 1 \leq i_1 < \cdots < i_{|u|} \leq N$ is a sub-vector of $\boldsymbol{X}$ and $\boldsymbol{X}_{-u} = \boldsymbol{X}_{\{1, \ldots, N\} \backslash u}$ is the complement of $\boldsymbol{X}_u$. The marginal density function is $f_u(\boldsymbol{x_u}) := \int f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x_{-u}}$ for a given set $\emptyset \neq u \subseteq \{1, \ldots N\}$. $y(\boldsymbol{X}) := y(X_1, \ldots, X_N)$ is a mathematical model with random variables as inputs. We write a vector space of square-integrable functions as

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

.

$\mathcal{L}^2(\Omega, \mathcal{F}, \nu)$ is a Hilbert space with the inner product defined as:

$$\langle y, g \rangle = \int y(\boldsymbol{x}) g(\boldsymbol{x}) \, f_{\boldsymbol{X}} d\nu(\boldsymbol{x}) = \mathbb{E}[y(\boldsymbol{X}) g(\boldsymbol{X})].$$

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

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

Orthogonal projection

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 $\mathcal{L}^2$:

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

Definition of $\mathcal{L}^2$ space and projection modified from Nagler (2024).

Unconditional and Conditional expectation

$\mathbb{E}[.]$ denotes the expectation operator, $Var[.] := \mathbb{E}[(. - \mathbb{E}[.])^2]$ denotes the variance and $Cov[., *] := \mathbb{E}[(. - \mathbb{E}[.])(* - \mathbb{E}[*])]$ denotes the covariance operator.

In general, we define the conditional expectation of a vector of random variables $\boldsymbol{X} = (X_1, X_2)$ as follows:

$$\mathbb{E}[g(X_1, X_2) \mid X_1 = x_1] = \int g(x_1, s_2)\, p_{X_2|X_1}(s_2 \mid x_1)\, ds_2.$$

Only when X_1 and X_2 are independent can we write

$$\mathbb{E}[g(X_1, X_2) \mid X_1 = x_1] = \int g(x_1, s_2)\, p_{X_2|X_1}(s_2 \mid x_1)\, ds_2 = \int g(x_1, s_2)\, p_{X_2}(s_2)\, ds_2 = \mathbb{E}_{X_2}[g(x_1, X_2)].$$

Extended to n random variables it looks as follows. Without loss of generality, we condition on $X_1 = x_1$:

$$\begin{aligned}
\mathbb{E}[g(X_1, \ldots, X_n) \mid X_1 = x_1] &= \int g(x_1, s_2, \ldots, s_n)\, p_{X_2,\ldots,X_n|X_1}(s_2, \ldots, s_n \mid x_1)\, ds_2 \ldots ds_n \\
&= \int g(x_1, s_2, \ldots, s_n)\, p_{X_2}(s_2, \ldots, s_n)\, ds_2 \ldots, ds_n \\
&= \mathbb{E}_{X_2,\ldots,X_n}[g(x_1, X_2, \ldots, X_n)]
\end{aligned}$$

Properties of the Multivariate Normal Distribution

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a d-dimensional multivariate normal (MVN) random vector, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is the symmetric positive semi-definite covariance matrix.

The marginal distribution of X_i is generally given by an univariate normal distribution:

$$X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii}) \quad \text{for all } i = 1, \ldots, d.$$

If we condition on a subset of the variables, we can also make statements about the conditional distribution. For this we partition the random vector $\boldsymbol{X}$ into two parts, $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$, where $\boldsymbol{X}_A$ contains the variables we condition on and $\boldsymbol{X}_B$ contains the remaining variables. The joint distribution of $\boldsymbol{X}$ can be expressed as:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_A \\ \boldsymbol{X}_B \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix} \right).$$

The conditional distribution of $\boldsymbol{X}_B$ given $\boldsymbol{X}_A = \boldsymbol{x}_A$ is

$$\boldsymbol{X}_B \mid \boldsymbol{X}_A = \boldsymbol{x}_A \sim \mathcal{N}\left(\boldsymbol{\mu}_B + \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}(\boldsymbol{x}_A - \boldsymbol{\mu}_A), \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{AB}\right).$$

For normally distributed random variables, we also know that $\mathrm{Cov}(X_i, X_j) = 0$, implies $X_i \perp X_j$.

3   Background and Related Work

The related literature on 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 orthogonal components of increasing dimensionality. Independently, Sobol (1993) proved that any square integrable function on the unit hypercube can be decomposed into a sum of orthogonal components. He originally called it "decomposition into summands of different dimension" and later renames it "ANOVA-representation" (Sobol, 2001) now referred to as the fANOVA decomposition. 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 approaches. 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. 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 variance partitioning and interpretation via Sobol indices. 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. WWhile 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 vibrant 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 interpretable ML. Here fANOVA can be used as a model-agnostic tool Hooker (2004), Fumagalli et al. (2025) are relevant or as foundational principle to build inherently interpretable models Hu et al. (2025). TThis is probably the most recent field of fANOVA in which research is actively ongoing.

Finally, there are specific domains of statistics, such as geostatistics, that explicitly build models on fANOVA framework (see Muehlenstaedt et al. (2012) for fANOVA Kriging models) or use it to study functions arising in computational finance Liu and Owen (2006).

4 Formalization of fANOVA

We start by defining the fANOVA decomposition in a very general form (which is independent of distribution assumptions or anything of the sort).

Definition 4.1. *Let y denote a mathematical model with input denoted by $X_1, \ldots, X_N$. The functional ANOVA (fANOVA) decomposition of y takes the form:*

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

where $u \subseteq \{1, \ldots, N\} = \{\{1\}, \{2\}, \{1, 2\}, \ldots, \{1, \ldots, N\}\}$ is the set, which contains all subsets of the indices $1, \ldots, N$.

The decomposition consists of 2^N terms and gives us a very general expression which's specific form is determined by the assumptions about the input variables and integration measure.

4.1 Classical fANOVA

For the classical case, originally proposed by Sobol (1993), we make the assumption of independent identically distributed (i.i.d.) input variables. This means we work with the measure space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \nu)$, and with a general measure ν defined on it. Under independence the joint probability density function (pdf) is given by the product over the marginal pdfs, i.e. $f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) = \prod_{i=1}^{N} f_{X_i}(x_i) \, d\nu(x_i)$, where $f_{X_i} : \mathbb{R} \to \mathbb{R}_0^+$ is the marginal probability density function of X_i defined on $(\Omega_i, \mathcal{F}_i, \nu_i)$ (or the previously defined measure space?).
Next, we formulate a condition, proposed by Rahman (2014), which we would like to hold for the fANOVA terms to be well-defined and interpretable.
The strong annihilating conditions require that the fANOVA terms integrate to zero w.r.t the individual variables contained in u and weighted by the individual marginal pdfs:

$$\int y_u(\boldsymbol{x}_u) f_{X_i}(x_i) \, d\nu(x_i) = 0, \quad \text{for } i \in u \neq \emptyset. \tag{2}$$

Proposition 4.1. *Given the strong annihilating conditions, the fANOVA components are centered around zero. The constant term is the only exception.*

$$\int y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) := \mathbb{E}[y_u(\boldsymbol{X}_u)] = 0 \tag{3}$$

Proof.

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

$$= \int_{\mathbb{R}^{|u|}} y_u(\boldsymbol{x_u}) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u) \, d\nu(\boldsymbol{x}_u)$$

$$= \int_{\mathbb{R}^{|u|}} y_u(\boldsymbol{x_u}) \prod_{i \in u} f_{X_i}(x_i) \, d\nu(\boldsymbol{x}_u)$$

$$= \int_{\mathbb{R}^{|u|-1}} \int_{\mathbb{R}} y_u(\boldsymbol{x_u}) f_{X_i}(x_i) \, dx_u \prod_{j \in u, j \neq i} f_{X_j}(x_j) = 0$$

$\square$

Proposition 4.2. *Given the strong annihilating conditions, it follows that the fANOVA terms are orthogonal to each other. If two sets of indices are not completely equivalent, i.e. $\emptyset \neq u \subseteq \{1, \ldots, N\}, \emptyset \neq v \subseteq \{1, \ldots, N\}$, and $u \neq v$, then it holds that:*

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

Proof.

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

$$= \int_{\mathbb{R}^{\mathbb{N}}} y_u(\boldsymbol{x_u}) y_v(\boldsymbol{x_v}) \prod_{i=1}^{N} f_{X_i}(x_i) \, d\nu(x_i)$$

$$= \int_{\mathbb{R}^{\mathbb{N}-1}} \int_{\mathbb{R}} y_u(\boldsymbol{x_u}) y_v(\boldsymbol{x_v}) f_{X_i}(x_i) \, dx_u \prod_{j \in \{1, \ldots, N\}, j \neq i} f_{X_j}(x_j) = 0$$

$\square$

This means that fANOVA terms are "fully orthogonal" to each other, meaning not only terms of different order are orthogonal to each other but also terms of the same order are. Zero-mean and orthogonality are desirable and important properties because they ensure that the fANOVA terms can be interpretate as isolated effects of the specific variable(s). The term y_1, for example, captures the isolated main effects of X_1; there is no other effect mixed into it, which X_1 might have through interactions with other variables. From the lense of interpretability, this distinguishes the fANOVA decomposition from methods such as partial dependence (PD) or Shapley values.

Construction of the fANOVA Terms

The individual fANOVA terms for the variables with indices in u are constructed by integrating the original function $y(\boldsymbol{X})$ w.r.t all variables expect for the ones in u, and subtracting the lower order terms. Intuitively the integral is averaging the original function over all other variables expect the ones of interest, which makes sense as we are then left with a function of the variables of interest only. Subtracting lower order terms corresponds to account for effects that are already explained by other variables or interactions so that we obtain the isolated effects.

Since $u = \emptyset$ for the constant term, we integrate w.r.t all variables:

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

For all other effects $\emptyset \neq u \in \{1, \ldots, N\}$ we can calculate:

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

Notice that this definition relies on a product-type measure rooted in the independence assumption. We will see what changes when we let got of this assumption in the next section.

As suggested earlier, the fANOVA components 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 and XAI literature, holding the potential for a global model-agnostic explanation method of black box models.

4.1.1 Example: Multivariate Normal Inputs

Before further investigating the fANOVA decomposition, let us consider the following function as example: $g = a + X_1 + 2X_2 + X_1 X_2$. We assume that $\boldsymbol{X} = (X_1, X_2)^T$ follows a standard MVN distribution, so the $\boldsymbol{\mu} = (0, 0)^T$ and the covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

From the properties of the MVN, we know that marginal distributions are standard normal:

$$X_i \sim \mathcal{N}(0, 1) \quad \text{for } i = 1, 2$$

We also know that the conditional distributions are given by:

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}(0, 1), \quad X_2 \mid X_1 = x_1 \sim \mathcal{N}(0, 1)$$

Case 1: Independent Inputs

The classical fANOVA decomposition we covered so far assumes $\rho_{12} = 0$. Computing the fANOVA decomposition of $g(x_1, x_2)$ by hand, we start with the constant term and make use of formulation via the expected value:

$$y_0 = \mathbb{E}[g_1(X_1, X_2)] = \mathbb{E}[a + X_1 + 2X_2 + X_1X_2] = \mathbb{E}[a] + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1X_2]$$

Making use of the independence assumption of X_1 and X_2, the last term can be written as the product of the expected values. Additionally, given the zero-mean property, all terms, except for the constant, vanish, and we obtain:

$$y_0 = \mathbb{E}[a] + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1]\mathbb{E}[X_2] = a$$

Under zero-mean constraint and independence, the main effects and the interaction effect can be computed as follows:

$$
\begin{aligned}
y_1(x_1) &= \mathbb{E}_{X_2}[g_1(x_1, X_2)] - y_0 \\
&= \mathbb{E}_{X_2}[a + x_1 + 2X_2 + x_1X_2] - a \\
&= x_1 + 2\mathbb{E}[X_2] + x_1\mathbb{E}[X_2] = x_1 \\
y_2(x_2) &= \mathbb{E}_{X_1}[g_1(X_1, x_2)] - y_0 \\
&= \mathbb{E}_{X_1}[a + X_1 + 2x_2 + X_1x_2] - a \\
&= \mathbb{E}_{X_1}[X_1] + 2x_2 + x_2\mathbb{E}_{X_1}[X_1] = 2x_2 \\
y_{12}(x_1, x_2) &= \mathbb{E}[g_1(x_1, x_2)] - y_0 - y_1(x_1) - y_2(x_2) \\
&= a + x_1 + 2x_2 + x_1x_2 - a - x_1 - 2x_2 = x_1x_2
\end{aligned}
$$

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.

4.1.2 fANOVA as projection

In the following we revisit the fANOVA decomposition from the view of orthogonal projections. The section is based on Van der Vaart (1998). Having this perspective on the fANOVA decomposition is useful 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 in section **??**.

When we define the constant term $y_\emptyset$ 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 $g(x) = a$ over all possible constant functions. The solution is the orthogonal projection of y onto the linear subspace of all constant functions $\mathcal{G}_0 = \{g(x) = a; a \in \mathbb{R}\}$. In a probabilistic context, we want to minimize the expected squared different between the random variables $y(\boldsymbol{X})$ and a, which turns out to be equivalent to the expected value of the random variable (Van der Vaart, 1998). So intuitively, in the absence of any additional information, the expected value is our best approximation of y. More formally we can write:

$$\Pi_{\mathcal{G}_0} y = \arg \min_{g_0 \in \mathcal{G}_0} \|y - g_0\|^2$$
$$= \arg \min_{a_0 \in \mathbb{R}} \mathbb{E}[(y(\boldsymbol{X}) - a)^2]$$
$$= \mathbb{E}[y(\boldsymbol{X})] = y_0$$

The main effect $y_i(x_i)$ is the projection of y onto the subspace of all functions that only depend on x_i, i.e. $\mathcal{G}_i = \{g(x) = g_i(x_i)\}$. There is no need for additional constraints since subtracting lower order terms ensures that orthogonality and zero mean are fulfilled. The conditional expected value of $\mathbb{E}[y(\boldsymbol{X}) \mid X_i = x_i]$ is the solution to the minimization problem (Van der Vaart, 1998), and the conditional expected value is also a way to express the fANOVA terms (Muehlenstaedt et al., 2012):

$$(\Pi_{\mathcal{G}_i} y)(.) - y_0 = \arg \min_{g_i \in \mathcal{G}_i} \|y - g_i\|^2 - y_0$$
$$= \arg \min_{g_i \in \mathcal{G}_i} \mathbb{E}[(y(\boldsymbol{X}) - g_i(X_i))^2] - y_0$$
$$= \mathbb{E}[y(\boldsymbol{X}) \mid X_i = .] - y_0 = y_i(.)$$

The two-way interaction effect $y_{ij}(.,.)$ is the projection of y onto the subspace of all functions that depend on x_i and x_j. i.e. $\mathcal{G}_{i,j} = \{g(x) = g_{ij}(x_i, x_j)\}$. Again, we account for lower-order effects by subtracting the constant term and all main effects:

$$(\Pi_{\mathcal{G}_{ij}} y)(.;.) - (y_0 + y_i(.) + y_j(.)) = \arg \min_{g_{ij} \in \mathcal{G}_{ij}} \|y - g_{i,j}\|^2 - (y_0 + y_i(.) + y_j(.))$$
$$= \arg \min_{g_{ij} \in \mathcal{G}_{ij}} \mathbb{E}[(y(\boldsymbol{X}) - g(.,.))^2] - (y_0 + y_i(.) + y_j(.))$$
$$= \mathbb{E}[y(\boldsymbol{X})|X_j = x_j, X_i = x_i] - (y_0 + y_i(.) + y_j(.)) = y_{ij}(.;.)$$

In general, we can write for a subset of indices $u \subseteq \{1, \dots, N\}$ and the subspace $\mathcal{G}_u =$

$\{g(\boldsymbol{x}) = g_u(\boldsymbol{x}_u)\}$:

$$(\Pi_{\mathcal{G}_u} y)(.) - \sum_{v \subsetneq u} y_v(.) = \arg \min_{g_u \in \mathcal{G}_u} \|y - g_u\|^2 - \sum_{v \subsetneq u} y_v(.)$$

$$= \arg \min_{g_u \in \mathcal{G}_u} \mathbb{E}[(y(\boldsymbol{X}) - g(.))^2] - \sum_{v \subsetneq u} y_v(.)$$

$$= \mathbb{E}[y(\boldsymbol{X}) | X_u = x_u] - \sum_{v \subsetneq u} y_v(x) = y_u(.),$$

which means that we project y onto the subspace spanned by the own terms of the fANOVA component to be defined, while accounting for all lower-order terms.

Projection of the differences or subtracting from the projection

Thanks to the equivalence of the conditional expected value and projections we established the mathematical foundation/ mechanism of fANOVA. Next we want to highlight that instead of subtracting the lower order terms from the projection, it is just as valid to first subtract lower order terms 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:

$$y_u(\boldsymbol{x}_u) := \mathbb{E}[y(\boldsymbol{X}) | \boldsymbol{X}_u = \boldsymbol{x}_u] - \sum_{v \subsetneq u} y_v(\boldsymbol{x})$$

$$\int_{-u} y(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-u}) - \sum_{v \subsetneq u} y_v(\boldsymbol{x})$$

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

$$y_u(\boldsymbol{x}_u) := \int_{-u} (y(\boldsymbol{x}) - \sum_{v \subsetneq u} y_v(\boldsymbol{x})) d\nu(\boldsymbol{x}_{-u})$$

$$\mathbb{E}[y(\boldsymbol{X}) - \sum_{v \subsetneq u} y_v(\boldsymbol{x}) | \boldsymbol{X}_u = \boldsymbol{x}_u]$$

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.

Second-moment statistics

No handbook on fANOVA is complete without at least mentioning *Sobol indices*. This requires us to observe the second moment statistics of the decomposition. We already

established that $\mathbb{E}[y(\boldsymbol{X})] = y_\emptyset$. We can also compute the variance of $y(\boldsymbol{X})$ via the fANOVA decomposition. The variance is defined as the expected value of the squared difference between the random variable and its expected value:

We write the sum over u for the sum over $\emptyset \neq u \subseteq \{1, \ldots, N\}$ and the sum over $u \neq v$ for the sum over $\emptyset \neq u \subseteq \{1, \ldots, N\}, \emptyset \neq v \subseteq \{1, \ldots, N\}, u \neq v$.

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

We can verify that the variance decomposition holds for our example:

$$
\begin{aligned}
Var(a + X_1 + 2X_2 + X_1 X_2) &= Var(X_1) + 4Var(X_2) + Var(X_1 X_2) + 2Cov(X_1, X_2) \\
&= 1 + 4 \cdot 1 + 1 \cdot 1 + 2 \cdot 0 = 6 \\
&= \mathbb{E}[X_1^2] + 4\mathbb{E}[X_2^2] + \mathbb{E}[X_1^2]\mathbb{E}[X_2^2] + 2Cov(X_1, X_2) \\
&= \mathbb{E}[y_1^2] + \mathbb{E}[y_2^2] + \mathbb{E}[y_{12}^2]
\end{aligned}
$$

Studying the variance of the decomposition was the main focus in early works on this method (see e.g. Sobol (1993)). From the variance decomposition Sobol (1993) construct the *Sobol indices*, which are well-known in sensitivity analysis. As it is only one application of the fANOVA decomposition, we will not go into depth here, but we should keep in mind that in most works, the presentation of fANOVA is closely linked to the Sobol indices.

Motivating Example

Recall our example setup of standard MVN input variables and $g = a + X_1 + 2X_2 + X_1 X_2$ from the previous section **??**. For classical fANOVA we make the assumption of independent inputs, which is often violated in practice. Let us therefore investigate what happens, when we allow for dependency between variables.

Case 2: Dependent Inputs

Now $\rho_{12} \neq 0$, while keeping everything else the same. When we follow the exact same logic as above we obtain the following terms:

$$\tilde{y}_0 = \mathbb{E}[g(X_1, X_2)] = a + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1 X_2]$$
$$= a + \mathbb{E}[X_1 X_2] = a + (\text{Cov}(X_1, X_2) + \mathbb{E}[X_1]\mathbb{E}[X_2])$$
$$= a + \rho_{12}$$
$$\tilde{y}_1(x_1) = \mathbb{E}[g(X_1, X_2)|X_1 = x_1] - \tilde{y}_0$$
$$= \mathbb{E}[a + x_1 + 2X_2 + x_1 X_2|X_1 = x_1] - (a + \rho_{12})$$
$$= a + x_1 + 2\mathbb{E}[X_2|X_1 = x_1] + x_1 \mathbb{E}[X_2|X_1 = x_1] - a - \rho_{12}$$
$$= x_1 + \rho_{12}(2x_1 + x_1^2 - 1)$$
$$\tilde{y}_2(x_2) = \mathbb{E}[g(X_1, X_2) \mid X_2 = x_2] - \tilde{y}_0$$
$$= \mathbb{E}[a + X_1 + 2x_2 + X_1 x_2 \mid X_2 = x_2] - (a + \rho_{12})$$
$$= a + 2x_2 + x_2 \mathbb{E}[X_1 \mid X_2 = x_2] - a - \rho_{12}$$
$$= 2x_2 + \rho_{12}(x_2 + x_2^2 - 1)$$
$$\tilde{y}_{12}(x_1, x_2) = g(x_1, x_2) - \tilde{y}_0 - \tilde{y}_1(x_1) - \tilde{y}_2(x_2)$$
$$= a + x_1 + 2x_2 + x_1 x_2 - (a + \rho_{12})$$
$$- (x_1 + \rho_{12}(2x_1 + x_1^2 - 1)) - (2x_2 + \rho_{12}(x_2 + x_2^2 - 1))$$
$$= x_1 x_2 - 2\rho_{12} x_1 - \rho_{12} x_2 - \rho_{12} x_1^2 - \rho_{12} x_2^2 + \rho_{12}$$

The fANOVA components are characterized by two central properties zero mean and orthogonality which follow from Equation 2. When we check if the components $\tilde{y}_0, \tilde{y}_1, \tilde{y}_2, \tilde{y}_{12}$ satisfy these properties, we find out that all components are zero-centred, but not all are orthogonal to each other. We can, for example, immediately see that checking orthogonality between $\tilde{y}_1, \tilde{y}_{1,2}$ will yield the expectation over the constant term $\rho_{1,2}$ exactly once, meaning even if all the other expectations cancel out, this constant will remain and the

entire expression will be unequal to zero:

$$\mathbb{E}(\tilde{y}_1(X_1)\tilde{y}_{1,2}(X_1, X_2)) = \mathbb{E}[(X_1 + 2\rho_{12}X_1 + \rho_{12}X_1^2 - \rho_{12})$$
$$\cdot (X_1X_2 - 2\rho_{12}X_1 - \rho_{12}X_2 - \rho_{12}X_1^2 - \rho_{12}X_2^2 + \rho_{12})]$$
$$= \mathbb{E}[X_1^2 X_2] \ldots - \mathbb{E}[\rho_{12}^2] \neq 0.$$

When we no longer have independent inputs naively computing the "fANOVA decomposition" does not yield the fANOVA components as it turns out. What we performed in this example is not the fANOVA decomposition for dependent variables. It is Hoeffding decomposition (Hoeffding, 1948) and results in zero mean but not mutually orthogonal component functions. This shows the need for a more involved approach for generalizing fANOVA. We basically can see from this example that correlation between features distorts the fANOVA component function, it is not pure anymore but this is a crucial point about fANOVA for interpretability.

4.2   Generalized fANOVA

We base this chapter mainly on the generalization of Rahman (2014), while there exists other work from Hooker (2007) or Chastaing et al. (2012). (Write this a bit more detailed: Hooker (2007) proofed existence of generalized fANOVA components, proposed estimation scheme, Rahman (2014) writes this in more general and measure theoretic fashion and proposes different estimation scheme that he argues is more feasible for high dimensions etc. read more in intro of Rahman (2014); Hooker (2007) seems to be viewed as the first one who attempted a generalization to dependent inputs of the entire fANOVA decomposition framework, not just the Sobol indices, and he was inspired by Stone (1994)). Letting go of the independence assumption means that we no longer work with a product-type probability measure. $f_{\boldsymbol{X}} : \mathbb{R} \to \mathbb{R}_0^+$ denotes an arbitrary probability density function and $f_{\boldsymbol{X}_u} : \mathbb{R}^u \to \mathbb{R}_0^+$ the marginal probability density function of the subset of variables $u \subseteq d$. Classical fANOVA boils down to integration w.r.t. the uniform measure and in generalized fANOVA we integrate w.r.t. the distribution of $(X_1, \ldots, X_n)$.

Other than that, the generalized fANOVA decomposition still follows the overarching form from the very beginning Equation 1.

Instead of enforcing the strong annihilating conditions for desirable properties of the components, Rahman (2014) proposed to formulate a milder version. The milder version fulfills the same function as the strong annihilating conditions in the classical case but works with the joint density of the variables of interest, instead of the individual marginal probability density functions. **The weak annihilating conditions** require that for the

fANOVA component of variables in u integrate to zero w.r.t. the joint pdf of variables in u:

$$\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u) d\nu(x_i) = 0 \quad \text{for} \quad i \in u \neq \emptyset \tag{7}$$

If components are constructed in this way, we can ensure that they have zero mean and satisfy a milder from 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 4.3. *Given the weak annihilating conditions, the generalized fANOVA components $y_{u,G}$, with $\emptyset \neq u \subseteq \{1, \dots, N\}$, are centred around zero:*

$$\mathbb{E}[y_{u,G}(\boldsymbol{X}_u)] := \int y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) = 0 \tag{8}$$

Proof. For any subset $\emptyset \neq u \subseteq \{1, \dots, N\}$, let $i \in u$. We assume that the weak annihilating conditions hold. Then

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

where we make use of Fubini's theorem and the last line follows from using the weak annihilating condition $\qquad\square$

Proposition 4.4. *Given the weak annihilating conditions, the fANOVA components are hierarchically orthogonal. This means that for two components $y_{u,G}$ and $y_{v,G}$ with $u \subsetneq v, \emptyset \neq u \subseteq \{1, \dots, N\}, \emptyset \neq v \subseteq \{1, \dots, N\}$ it holds that:*

$$\mathbb{E}[y_{u,G}(\boldsymbol{X}_u) y_{v,G}(\boldsymbol{X}_v)] := \int y_{u,G}(\boldsymbol{x}_u) y_{v,G}(\boldsymbol{X}_v) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) = 0 \tag{9}$$

Proof. For any two subsets $\emptyset \neq u \subseteq \{1, \dots, N\}$ and $\emptyset \neq v \subseteq \{1, \dots, N\}$, where $v \subsetneq u$,

17

the subset $u = v \cup (u \setminus v)$. Let $i \in (u \setminus v) \subseteq u$. Then

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

Repeatedly using Fubini's theorem and the weak annihilating conditions the equality to zero follows. $\qquad\square$

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 the Sobol indices. This means it is important that Rahman's generalized statements reduce to the classical case under product-type pdf.

Proposition 4.5. *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 $f_u(\mathbf{x}_u)$ as

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

Now consider the weak annihilating condition (4.2) for some $i \in u \neq \emptyset$:

$$
\int_{\mathbb{R}} y_{u,G}(\mathbf{x}_u) f_u(\mathbf{x}_u)\, dx_i = 0.
$$

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

$$
\int_{\mathbb{R}} y_{u,G}(\mathbf{x}_u) \left(\prod_{j \in u} f_{\{j\}}(x_j) \right) dx_i.
$$

For fixed x_j with $j \neq i$, the terms $f_{\{j\}}(x_j)$ are constant with respect to x_i, and can therefore be pulled out of the integral:

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

As product of pdfs the prefactor is strictly positive for all x_j with $j \neq i$. Therefore, the integral must be zero for the equality to hold:

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

which are the strong annihilating conditions from the previous section. $\square$

Construction of the Generalized fANOVA Terms

Recall the construction of the classical fANOVA components Equation 6. The equation tells us that the non-constant classical fANOVA components are defined via the integral of the original function w.r.t. to the product-type pdf, minus effects by other terms. So ideally for a well-aligned generalization, we would want that the general fANOVA terms can be understood in a similar way, as the integral of y w.r.t. *some type of pdf*, minus effects explained by other terms. This is exactly what Rahman (2014) accomplishes. To understand this, we first need to distinguish three cases of integration that will occure in the construction of the generalized components.

Proposition 4.6. *Consider the generalized fANOVA components* $y_{v,G}$, $\emptyset \neq v \subseteq \{1, \ldots, N\}$, *of a square-integrable function* $y : \mathbb{R}^N \to \mathbb{R}$. *When integrated w.r.t. the probability measure* $f_{-u}(\boldsymbol{x}_{-u}) \, d\boldsymbol{x}_{-u}$, $u \subseteq \{1, \ldots, N\}$, *one can distinguish three cases:*

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

Proof. Let $u \subseteq \{1, \ldots, N\}$ and $\emptyset \neq v \subseteq \{1, \ldots, N\}$. We distinguish between three types of relationship 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** $v \not\subset u$. We use the decomposition of $-u$ stated above to decompose the integration over $\mathbf{x}_{-u}$ as:

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

The inner integral gives the marginal density $f_{v \cap -u}(\mathbf{x}_{v \cap -u})$, so we obtain:

$$= \int_{\mathbb{R}^{|v \cap -u|}} y_{v,G}(\mathbf{x}_v) f_{v \cap -u}(\mathbf{x}_{v \cap -u}) \, d\mathbf{x}_{v \cap -u}.$$

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

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

which works because f_{-u} is a pdf.

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

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

$$= \int_{\mathbb{R}^{|v|}} y_{v,G}(\mathbf{x}_v) f_v(\mathbf{x}_v) \, d\mathbf{x}_v$$

$$= \int_{\mathbb{R}^{|v|-1}} \left(\int_{\mathbb{R}} y_{v,G}(\mathbf{x}_v) f_v(\mathbf{x}_v) \, dx_i \right) \prod_{\substack{j \in v \\ j \neq i}} dx_j$$

$$= 0,$$

while we again split the interval in such a way that we recognize the marginal density f_v and make use of the zero mean property from the strong annihilating conditions.

$\square$

As we will see in the following, we will encounter all of these three cases in the definition of the generalized fANOVA components via Rahman (2014) principle. It just remains to state the pdf w.r.t. which we integrate. Rahman proposes $f_{-u}(\boldsymbol{x}_{-u})$.

Theorem 4.1. *The generalized fANOVA component functions $y_{u,G}(\mathbf{x}_u)$ can be recursively*

defined via the following set of equations:

$$y_{\emptyset,G} = \int_{\mathbb{R}^N} y(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}, \tag{4.5a}$$

$$y_{u,G}(\mathbf{X}_u) = \int_{\mathbb{R}^{N-|u|}} y(\mathbf{X}_u, \mathbf{x}_{-u}) f_{-u}(\mathbf{x}_{-u}) \, d\mathbf{x}_{-u} - \sum_{v \subset u} y_{v,G}(\mathbf{X}_v)$$

$$- \sum_{\substack{\emptyset \neq v \subseteq \{1,\ldots,N\} \\ v \cap u \neq \emptyset, \; v \not\subset u}} \int_{\mathbb{R}^{|v \cap -u|}} y_{v,G}(\mathbf{X}_{v \cap u}, \mathbf{x}_{v \cap -u}) f_{v \cap -u}(\mathbf{x}_{v \cap -u}) \, d\mathbf{x}_{v \cap -u}. \tag{4.5b}$$

Proof. We begin by integrating both sides of the generalized fANOVA decomposition

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

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

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

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

$$\int_{\mathbb{R}^N} y(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x} = \sum_{v \subseteq \{1,\ldots,N\}} \int_{\mathbb{R}^N} y_{v,G}(\mathbf{x}_v) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}$$

$$= \int_{\mathbb{R}^N} y_{\emptyset,G} \, f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x} + \sum_{\emptyset \neq v \subseteq \{1,\ldots,N\}} \int_{\mathbb{R}^N} y_{v,G}(\mathbf{x}_v) f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}$$

$$= y_{\emptyset,G} + \sum_{\emptyset \neq v \subseteq \{1,\ldots,N\}} \mathbb{E}[y_{v,G}(\mathbf{X}_v)] = y_{\emptyset,G},$$

where the last sum vanishes under the weak annihilating condition.

Case $\emptyset \neq u \subseteq \{1,\ldots,N\}$: computing nonconstant terms. We return to the integrated decomposition

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

and apply Lemma 4.3 to evaluate each term in the sum according to the relationship between v and u:

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

This is Case 1 of Lemma 4.3. The integral becomes:

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

(B) $v \subsetneq u$:

This is Case 2 of Lemma 4.3. The integrals reduce to the component functions themselves:

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

(C) $v = u$:

Also part of Case 2 of Lemma 4.3. The integral becomes:

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

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

Case 3 of Lemma 4.3. These terms vanish:

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

Putting everything together:

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

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

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

As a last step, we only have to write $v = (v \cap u) \cup (v \cap -u)$ to obtain the expression of Theorem 5.1.

$\square$

4.2.1 Alternative Definition of Generalized fANOVA Components

Hooker (2007) approaches his generalization of the fANOVA decomposition differently, from the angle of orthogonal projections. Instead of a more recursive definition of the components functions as in Rahman (2014), he defines the fANOVA components as a joint set which simultaneously minimizes the squared difference to the original function y under certain constraints. The constraints he sets for the optimization problem should ensure that the generalized components satisfy the desired properties of zero mean and hierarchical orthogonality.

The generalized fANOVA terms $\{y_u(x_u) | u \subseteq d\}$ jointly satisfiy:

$$\{y_{u,G}(\boldsymbol{x}_u) \mid u \subseteq d\} = \arg \min_{\{g_u \in L^2(\mathbb{R}^u)\}_{u \subseteq d}} \int \left(\sum_{u \subseteq d} g_u(\boldsymbol{x}_u) - y(\boldsymbol{x}) \right)^2 f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) \qquad (10)$$

under the hierarchical orthogonality conditions:

$$\forall v \subseteq u, \ \forall g_v : \int y_u(x_u) g_v(x_v) w(x) \, dx = 0. \qquad (4.2)$$

In Hookers definition we recognize a projection. We are simultaneously finding the set of components functions g_u that minimize the weighted squared difference to the original function y (under zero mean and hierarchical orthogonality constraint), which is exactly the definition of a projection of y onto a specific subspace $\mathcal{G}$, which we defined generally in section **??**.

The following proposition provides the foundation for Hookers generalization because it gives a tangible constraint, one can set to ensure that one really obtains fANOVA components (which satisfy hierarchical orthogonality) instead of the components of the Hoeffding decomposition we saw in our example. This proposition fulfills the role of the weak annihilating conditions in Rahman (2014).

Proposition 4.7. *The hierarchical orthogonality of the fANOVA components is ensured if and only if the following integral condition holds:*

$$\forall u \subseteq N, \ \forall i \in u : \ \int y_u(x_u) w(x) \, dx_i \, dx_{-u} = 0. \qquad (4.3)$$

Proof. The proof is organized in two parts. First, Hooker needs to show that, if the integral conditions hold, the hierarchical orthogonality is true, and second, that if the hierarchical orthogonality does not hold, the integral conditions do not hold either. For

the first part, assume that (4.3) holds. Let $i \in u \setminus v$, then $y_v(x_v)$ is independent of x_i and x_{-u}, so we can write:

$$\int y_v(x_v)\, y_u(x_u)\, w(x)\, dx_i\, dx_{-u} = y_v(x_v) \int y_u(x_u)\, w(x)\, dx_i\, dx_{-u} = 0. \tag{11}$$

For the second part, assume that there exists a subset u and an index i for which (4.3) does not hold, i.e.

$$\int y_u(x_u)\, w(x)\, dx_i\, dx_{-u} \neq 0. \tag{12}$$

Further, assume that (4.3) does hold for a subset $v \neq u$ and an index $j \in v$. Hooker then constructs a fANOVA term y_v with lower order than y_u, which is not orthogonal to y_u. He sets $v = u \setminus \{i\}$, so y_v is one order lower than y_u and defined as:

$$y_v(x_v) := \int f_u(x_u)\, w(x)\, dx_i\, dx_{-u}. \tag{13}$$

y_v is a valid fANOVA component, which is unequal to zero by assumption of (4.3) being false, while it itself satisfies (4.3):

$$\forall j \in v, \quad \int y_v(x_v)\, w(x)\, dx_j\, dx_{-v} = 0 \tag{14}$$

Lastly, Hooker verifies that f_v is not orthogonal to f_u:

$$\begin{aligned}
\langle y_u, y_v \rangle_w &= \int y_u(x_u)\, y_v(x_v)\, f_X(x)\, dx \\
&= \int y_u(x_u) \left(\int y_u(x_u)\, f_X(x)\, dx_i\, dx_{-u} \right) w(x)\, dx \\
&= \int \left(\int y_u(x_u)\, f_X(x)\, dx_i\, dx_{-u} \right)^2 dx_{u \setminus \{i\}} \\
&\neq 0.
\end{aligned} \tag{15}$$

$\square$

A crucial difference to the classical case is that both versions of the generalized components are defined in dependence of each other (**??**, Equation 10). In the Rahman approach, the components are derived from the original function by integrating out the other variables, while in the Hooker approach, the components are defined through a minimization problem that seeks to best approximate the original function. This makes it in general difficult to compute the generalized fANOVA components analytically, even for simple functions.

Second-moment statistics

As mentioned earlier Sobol indices and the variance decomposition of the fANOVA terms is a curcial application of this method. While this work does not focus on Sobol indices, we note that they are based on the variance decomposition, which for the generalized case is also possible, but slightly modified compared to the classical case.

$$
\begin{aligned}
\sigma^2 &:= \mathbb{E}\left[(y(\mathbf{X}) - \mu)^2\right] \\
&= \mathbb{E}\left[\left(y_{\emptyset,G} + \sum_{\emptyset \neq u \subseteq \{1,\ldots,N\}} y_{u,G}(\mathbf{X}_u) - \mu\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{\emptyset \neq u \subseteq \{1,\ldots,N\}} y_{u,G}(\mathbf{X}_u)\right)^2\right] \\
&= \sum_{\emptyset \neq u} \mathbb{E}\left[y_{u,G}^2(\mathbf{X}_u)\right] + \sum_{\substack{\emptyset \neq u, \, v \subseteq \{1,\ldots,N\} \\ u \neq v, \, u \not\subset v, \, v \not\subset u}} \mathbb{E}\left[y_{u,G}(\mathbf{X}_u) y_{v,G}(\mathbf{X}_v)\right].
\end{aligned} \tag{6.2}
$$

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. For the classical fANOVA decomposition, the second term is zero for any relationship between u and v, and we are left with only the sum of the individual variances.

4.2.2 Correction of the Example: Dependent Multivariate Normal Inputs

Given the "naive" approach of computing the generalized fANOVA components from earlier, which left us with the Hoeffding decomposition instead, it remains to answer how the *fANOVA* decomposition looks like. In general it is difficult arrive at an analytical solution because of the interdependence of the components. For the running example, which is a two-degree polynomial, Rahman (2014) provides a way to obtain the fANOVA decomposition under dependent inputs.

The idea in Rahman (2014) is based on Fourier-Polynomial expansion, which allows to write each generalized fANOVA component as a weighted sum of basis functions. The problem shifts from finding the fANOVA component functions to finding the basis functions which allow us to express the fANOVA terms. Rahman chooses Hermite polynomials as basis functions, which are by construction zero mean and hierarchical orthogonal. Satisfying these properties, the Hermite polynomial basis functions ensure fANOVA com-

ponents that are also zero mean and hierarchical orthogonal. The challenge that remains is to find the weights for the basis functions, which can be done via coefficient matching; at least for a polynomial of degree two.

A polynomial of degree two has the general form $y(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_{11} x_1^2 + a_{22} x_2^2 + a_{12} x_1 x_2$. We know that any such polynomial may be expressed as a sum of weighted basis functions Nagler (2024) of the form:

$$y(x_1, x_2) = c_0 + c_{1,1}\, \psi_{1,1}(x_1) + c_{2,1}\, \psi_{2,1}(x_2) + c_{1,2}\, \psi_{1,2}(x_1) + c_{2,2}\, \psi_{2,2}(x_2) + c\, \psi(x_1, x_2)$$

where the $\psi_{i,j}$ are the basis functions with corresponding weights $c_0, c_{1,1}, c_{2,1}, c_{1,2}, c_{2,2}, c_{12,11} \in \mathbb{R}$.

The idea is to carefully construct a set of basis functions which fulfill zero mean and hierarchical orthogonality. Then the expansion in these basis functions is already the fANOVA decomposition of a quadratic polynomial, i.e.

$$
\begin{aligned}
y(x_1, x_2) &= a_0 + a_1 x_1 + a_2 x_2 + a_{11} x_1^2 + a_{22} x_2^2 + a_{12} x_1 x_2 \\
&= c_0 + c_{1,1}\, \psi_{1,1}(x_1) + c_{2,1}\, \psi_{2,1}(x_2) + c_{1,2}\, \psi_{1,2}(x_1) + c_{2,2}\, \psi_{2,2}(x_2) + c\, \psi(x_1, x_2) \\
&= \underbrace{c_0}_{y_0} + \underbrace{\left(c_{1,1}\, \psi_{1,1}(x_1) + c_{1,2}\, \psi_{1,2}(x_1)\right)}_{y_1(x_1)} + \underbrace{\left(c_{2,1}\, \psi_{2,1}(x_2) + c_{2,2}\, \psi_{2,2}(x_2)\right)}_{y_2(x_2)} + \underbrace{c\, \psi(x_1, x_2)}_{y_{12}(x_1, x_2)}.
\end{aligned}
$$

Coming from the probability density of a multivariate normal distribution, Rahman (2014) chooses multivariate Hermite polynomials. We use a slightly simplified (unscaled) version of the proposed basis functions to find an explicit solution for our running example. The basis functions we work with are:

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

where ρ is the correlation coefficient between x_1 and x_2. So this formula will work for

dependent as well independent inputs.

What remains it to find the coefficients $c_0, c_{1,1}, c_{2,1}, c_{1,2}, c_{2,2}, c_{12,11}$ such the weighted sum of the basis really recovers the original polynomial. To find the correct weights, we plug in the basis functions and rearrange terms to recognize the groups more easily:

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

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:

$$
-c_{12,11} = a_{12} \qquad \Rightarrow \qquad c_{12,11} = -a_{12}
$$

$$
c_{1,2} + c_{12,11}\frac{\rho}{1+\rho^2} = a_{11} \qquad \Rightarrow \qquad c_{1,2} = a_{11} + \frac{\rho}{1+\rho^2}a_{12}
$$

$$
c_{2,2} + c_{12,11}\frac{\rho}{1+\rho^2} = a_{22} \qquad \Rightarrow \qquad c_{2,2} = a_{22} + \frac{\rho}{1+\rho^2}a_{12}
$$

$$
c_{1,1} = a_1
$$

$$
c_{2,1} = a_2
$$

$$
c_0 - c_{1,2} - c_{2,2} + c_{12,11}\frac{\rho(\rho^2-1)}{1+\rho^2} = a_0 \qquad \Rightarrow \qquad c_0 = a_0 + a_{11} + a_{22} + \rho\,a_{12}
$$

So the (generalized) fANOVA decomposition of a two-degree polynomial is given by:

$$y(x_1, x_2) = c_0 + c_{1,1}\,\psi_{1,1}(x_1) + c_{2,1}\,\psi_{2,1}(x_2) + c_{1,2}\,\psi_{1,2}(x_1) + c_{2,2}\,\psi_{2,2}(x_2) + c\,\psi(x_1, x_2)$$

$$= \underbrace{(a_0 + a_{11} + a_{22} + \rho\,a_{12})}_{c_0} + \underbrace{a_1}_{c_{1,1}}\,x_1 + \underbrace{a_2}_{c_{2,1}}\,x_2$$

$$+ \underbrace{\left(a_{11} + \frac{\rho}{1+\rho^2}a_{12}\right)}_{c_{1,2}}(x_1^2 - 1) + \underbrace{\left(a_{22} + \frac{\rho}{1+\rho^2}a_{12}\right)}_{c_{2,2}}(x_2^2 - 1)$$

$$+ \underbrace{(-a_{12})}_{c_{12,11}}\left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2}\right)$$

$$= (a_0 + a_{11} + a_{22} + \rho a_{12}) + a_1 x_1 + a_2 x_2$$

$$+ \left(a_{11} + \frac{\rho}{1+\rho^2}a_{12}\right)(x_1^2 - 1) + \left(a_{22} + \frac{\rho}{1+\rho^2}a_{12}\right)(x_2^2 - 1)$$

$$- a_{12}\left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2}\right),$$

with individual components

$$y_0 = a_0 + a_{11} + a_{22} + \rho\,a_{12},$$

$$y_1(x_1) = a_1\,x_1 + \left(a_{11} + \frac{\rho}{1+\rho^2}a_{12}\right)(x_1^2 - 1),$$

$$y_2(x_2) = a_2\,x_2 + \left(a_{22} + \frac{\rho}{1+\rho^2}a_{12}\right)(x_2^2 - 1),$$

$$y_{12}(x_1, x_2) = -a_{12}\left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2}\right)$$

(16)

We are finally able to give the correct fANOVA terms for our running example in a generalized from, which allows for dependent inputs variables, assumed to be Gaussian. For $g(x_1, x_2) = x_1 + 2x_2 + x_1 x_2$ we have $a_0 = 0, a_1 = 1, a_2 = 2, a_{11} = 0, a_{22} = 0, a_{12} = 1$

and therefore obtain:

$$y_0 = \rho,$$

$$y_1(x_1) = x_1 + \frac{\rho}{1 + \rho^2}(x_1^2 - 1),$$

$$y_2(x_2) = 2\,x_2 + \frac{\rho}{1 + \rho^2}(x_2^2 - 1),$$

$$y_{12}(x_1, x_2) = -\left(\frac{\rho(x_1^2 + x_2^2)}{1 + \rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1 + \rho^2}\right).$$

5 Visualization and Estimation

5.1 Comparison of Decompositions

Time for some plots! Running example: $g(x_1, x_2) = x_1 + 2x_2 + x_1x_2$, with polynomial coefficients: $a_0 = 0$, $a_1 = 1$, $a_2 = 2$, $a_{11} = 0$, $a_{22} = 0$, $a_{12} = 1$ and $\rho = 0$. Under independent inputs, the fANOVA components are given by:

$$y_\emptyset = 0,$$
$$y_1(x_1) = x_1$$
$$y_2(x_2) = 2x_2$$
$$y_{12}(x_1, x_2) = x_1x_2,$$

visualized in Figure 1. One could vary the magnitude of the coefficients and obtain steeper, flatter, twisted (?) slopes for the main effects and also a steeper of flatter or twisted surface for the interaction term.
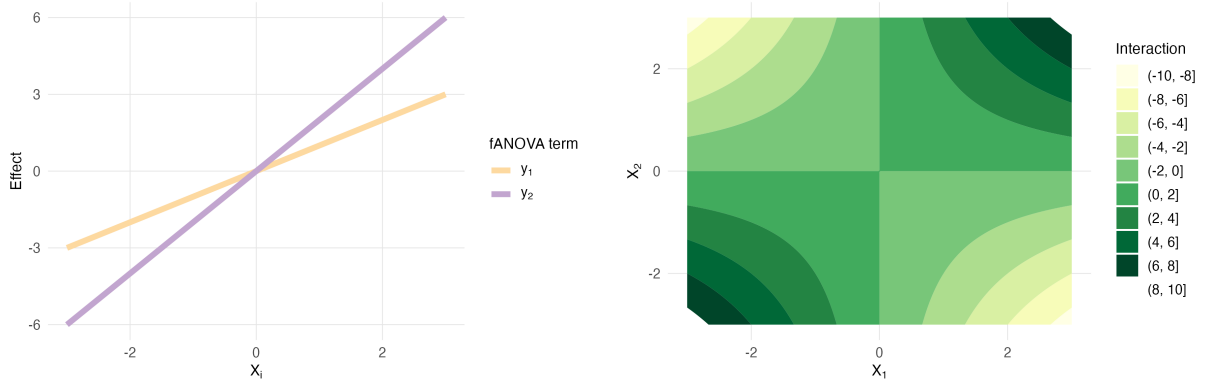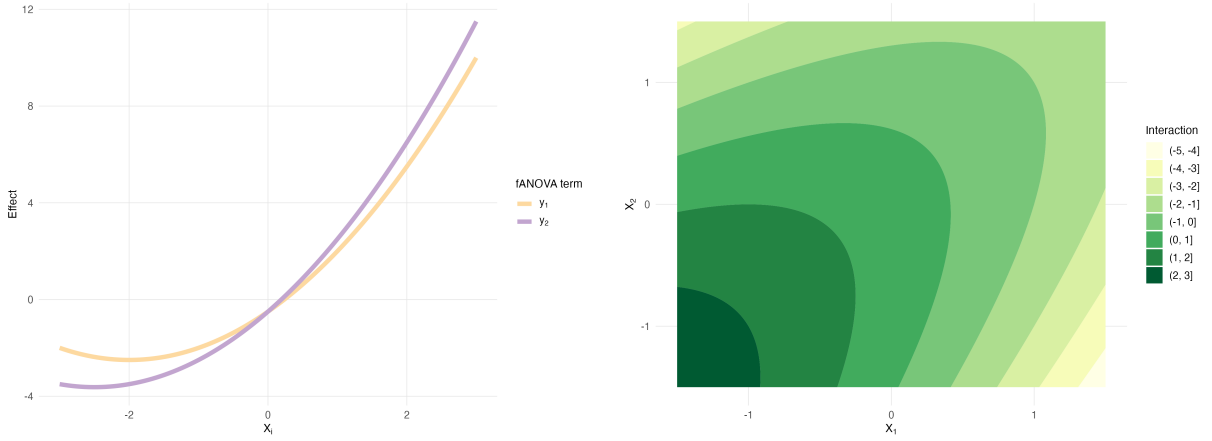

Figure 1: Main effects (left) and interaction effect (right) of the fANOVA decomposition for $g(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with independent inputs.

Running example: $g(x_1, x_2) = x_1 + 2x_2 + x_1x_2$, with polynomial coefficients: $a_0 = 0$, $a_1 = 1$, $a_2 = 2$, $a_{11} = 0$, $a_{22} = 0$, $a_{12} = 1$ and $\rho = 0.5$. In attempt to compute the fANOVA terms under dependent inputs, we calculated the following effects, which are in reality the components of the Hoeffding decomposition:

$$\tilde{y}_0 = a + 0.5$$

$$\tilde{y}_1(x_1) = x_1 + 0.5\left(2x_1 + x_1^2 - 1\right) = x_1 + x_1 + 0.5x_1^2 - 0.5 = 2x_1 + 0.5x_1^2 - 0.5$$

$$\tilde{y}_2(x_2) = 2x_2 + 0.5\left(x_2 + x_2^2 - 1\right) = 2x_2 + 0.5x_2 + 0.5x_2^2 - 0.5 = 2.5x_2 + 0.5x_2^2 - 0.5$$

$$\tilde{y}_{12}(x_1, x_2) = x_1x_2 - 2(0.5)x_1 - (0.5)x_2 - (0.5)x_1^2 - (0.5)x_2^2 + 0.5$$

$$= x_1x_2 - x_1 - 0.5x_2 - 0.5x_1^2 - 0.5x_2^2 + 0.5,$$

which are visualized in Figure 2. The true fANOVA components are given by:

Figure 2: Main effects (left) and interaction effect (right) of the Hoeffding decomposition for $g(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with dependent inputs, $\rho = 0.5$.

$$y_0 = 0.5$$

$$y_1(x_1) = x_1 + 0.4\left(x_1^2 - 1\right) = x_1 + 0.4x_1^2 - 0.4$$

$$y_2(x_2) = 2x_2 + 0.4\left(x_2^2 - 1\right) = 2x_2 + 0.4x_2^2 - 0.4$$

$$y_{12}(x_1, x_2) = -\left(0.4(x_1^2 + x_2^2) - x_1x_2 - 0.3\right)$$

$$= -0.4x_1^2 - 0.4x_2^2 + x_1x_2 + 0.3.$$

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

Our running example included linear effects of both input variables and the interaction term. For the remainder of this section, we will explore other representative scenarios, we can build within the scaffold of a bivariate two degree polynomial.
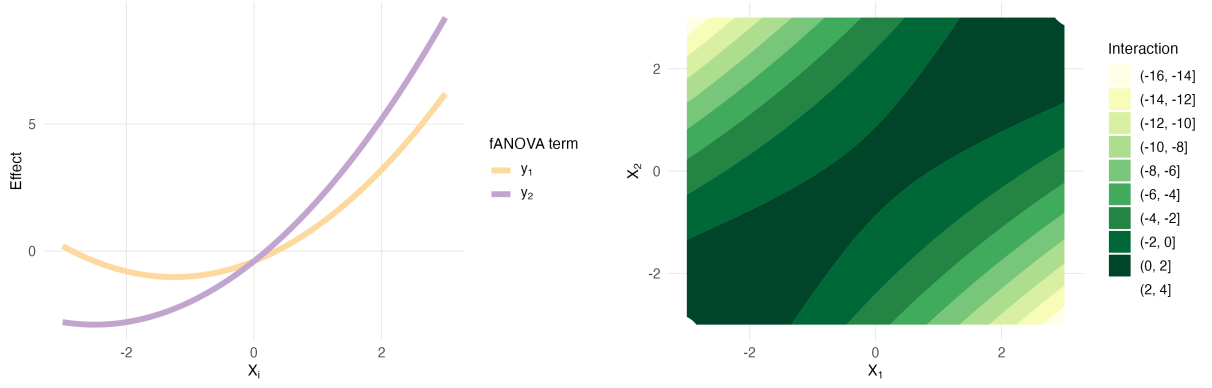
Figure 3: Main effects (left) and interaction effect (right) of the generalized fANOVA decomposition for $g(x_1, x_2) = x_1 + 2x_2 + x_1x_2$ with dependent inputs, $\rho = 0.5$.

5.2 Comparison of Functions

Scenario: Linear

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

$$g_1(x_1, x_2) = a_1x_1 + a_2x_2.$$

We can immediately read of the fANOVA components or see, that the general set of fANOVA components for a two degree polynomial in Equation 16 simplify for g_1 to:

$$y_1(x_1) = a_1x_1$$
$$y_2(x_2) = a_2x_2.$$

The sign of the coefficients a_1 and a_2 determines the direction of each slope respectively, while their magnitudes influence the steepness (Figure 4). Since an interaction effect is absent in g_1, the fANOVA interaction term is zero and omitted here; varying ρ or visualizing an interaction term is of no interest.

Scenario: Quadratic

Next we consider two-degree polynomials, which contain only quadratic terms. They take the form:
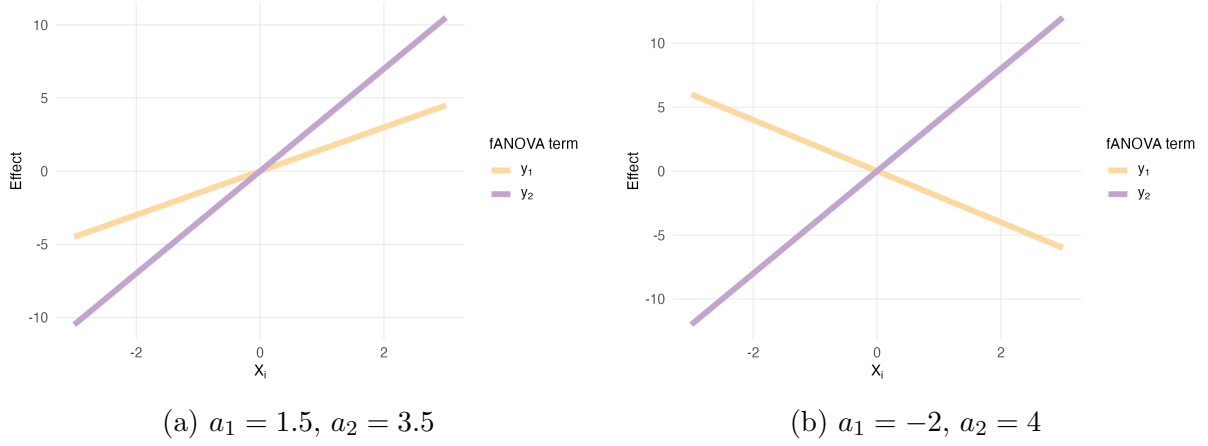
$$g_2(x_1, x_2) = a_{11}x_1^2 + a_{22}x_2^2.$$

(a) $a_1 = 1.5$, $a_2 = 3.5$ (b) $a_1 = -2$, $a_2 = 4$

Figure 4: Main fANOVA components for linear terms with different coefficients. The components are given by: $y_1(x_1) = a_1 x_1$, $y_2(x_2) = a_2 x_2$.

From Equation 16 we can derive the fANOVA components for g_2:

$$y_\emptyset = a_{11} + a_{22},$$
$$y_1(x_1) = a_{11}(x_1^2 - 1),$$
$$y_2(x_2) = a_{22}(x_2^2 - 1).$$

In Figure 5, we vary the coefficients a_{11} and a_{22}, while the interaction term is again absent. Intuitively, the sign of the coefficients determines the direction of the parabolas, while the magnitudes influence how stretched or compressed they are. The constant fANOVA component $y_\emptyset$ is present here, but as it only shifts the decomposition along the y-axis, it does not affect the form of the effect functions, therefore we do not visualize it.

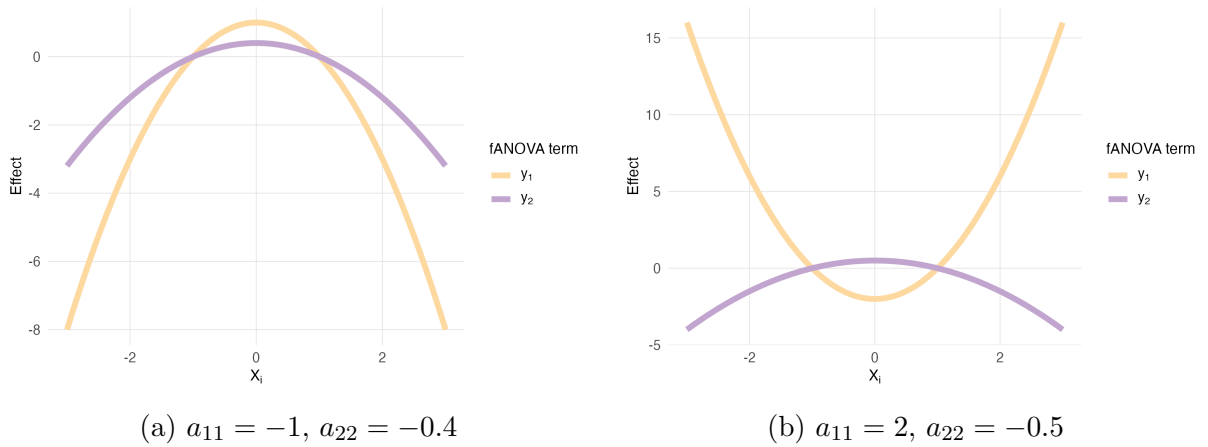(a) $a_{11} = -1$, $a_{22} = -0.4$ (b) $a_{11} = 2$, $a_{22} = -0.5$

Figure 5: Main effects for quadratic terms with different coefficients. The fANOVA components are given by: $y_1(x_1) = a_{11}(x_1^2 - 1)$, $y_2(x_2) = a_{22}(x_2^2 - 1)$.

Scenario: Interaction only

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

$$g_3(x_1, x_2) = a_{12}x_1x_2.$$

The fANOVA components for g_3 are given by:

$$y_\emptyset = a_{12}\rho,$$
$$y_1(x_1) = a_{12}\frac{\rho}{1+\rho}(x_1^2 - 1),$$
$$y_2(x_2) = a_{12}\frac{\rho}{1+\rho}(x_2^2 - 1),$$
$$y_{12}(x_1, x_2) = -a_{12}\left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2}\right).$$

What is interesting to see is that the main effects are present, due to the dependence between variable, even though there are no isolated main effects in g_3. The main effects y_1 and y_2, as well as the interaction term y_{12}, 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 Figure 6. For every case of ρ and a_{12}, the main effects have the same form, therefore they overlap.

Scenario: Mixed, only main

Following, we study the main effects of two-degree polynomials, which allow for effects of linear and quadratic nature:

$$g_4(x_1, x_2) = a_1x_1 + a_2x_2 + a_{11}x_1^2 + a_{22}x_2^2.$$

The fANOVA components for g_4 are given by:

$$y_\emptyset = a_{11} + a_{22},$$
$$y_1(x_1) = a_1x_1 + a_{11}(x_1^2 - 1),$$
$$y_2(x_2) = a_2x_2 + a_{22}(x_2^2 - 1).$$

In Figure 7, we vary the coefficients a_1, a_2, a_{11}, and a_{22}, while the interaction term is again absent. The shape of the parabolas in the upper left panel is representative of cases, where the coefficients of the linear terms a_1 and a_2 have the same sign and the ones of the quadratic terms a_{11} and a_{22} have the same sign. The coefficients of the quadratic
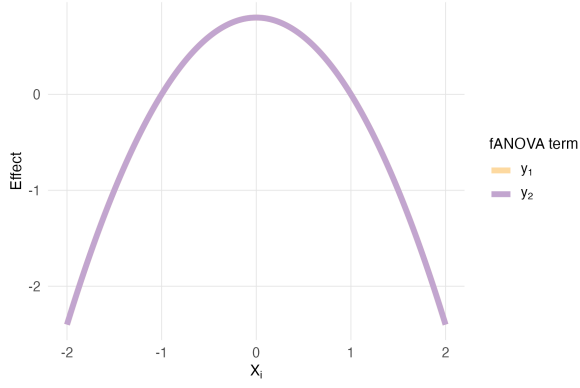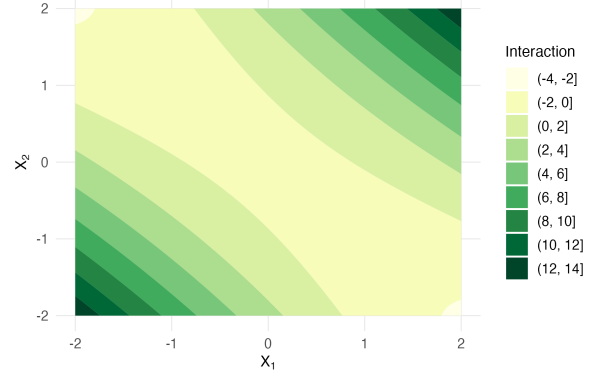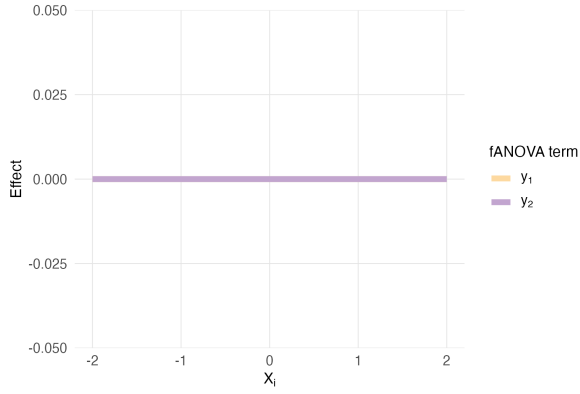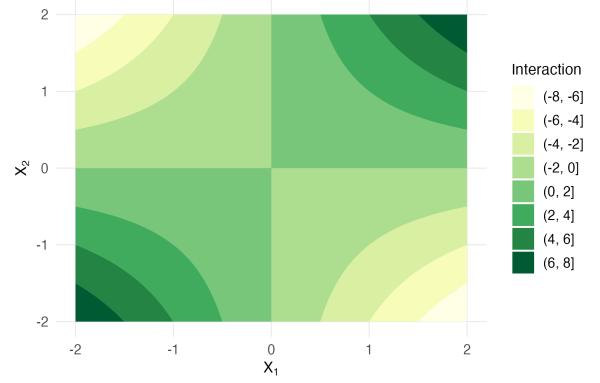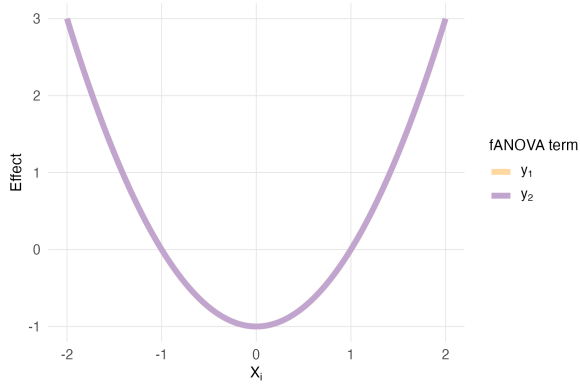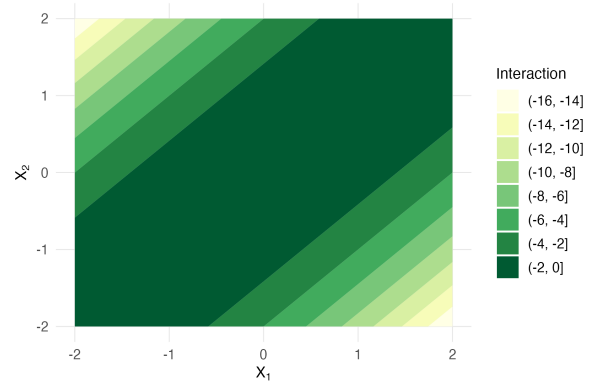
(a) Main effects for $\rho = -0.5$

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

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

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

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

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

Figure 6: Main effects (left column) and interaction contour plots (right column) for different values of ρ. The effects are given by $y_1(x_1) = a_{12}\frac{\rho}{1+\rho}(x_1^2 - 1)$, $y_2(x_2) = a_{12}\frac{\rho}{1+\rho}(x_2^2 - 1)$, $y_{12}(x_1, x_2) = -a_{12}\left(\frac{\rho(x_1^2 + x_2^2)}{1+\rho^2} - x_1 x_2 + \frac{\rho(\rho^2 - 1)}{1+\rho^2}\right)$.

terms a_{11} and a_{22} determine whether the parabola is facing downwards or upwards; when they are both negative the parabola is open downwards and when they are both positive the parabola is upwards open. Alongside the quadratic coefficients, the linear ones a_1 and a_2 influence how stretched or compressed the parabola is.

The upper right plot and the lower left plot show typical behaviour for cases where the quadratic coefficients have different signs. In such cases we get parabolas that are open in opposite directions. Lastly, in the lower right we have cases where the signs of the linear coefficient a_i is different than of the quadratic coefficient a_{ii} within the same Variable X_i, but the signs of the quadratic coefficients are the same between variables (so a_{ii} and a_{jj} have the same sign). Here we have a symmetric pattern, where the parabolas are open in the same direction and otherwise may look different.
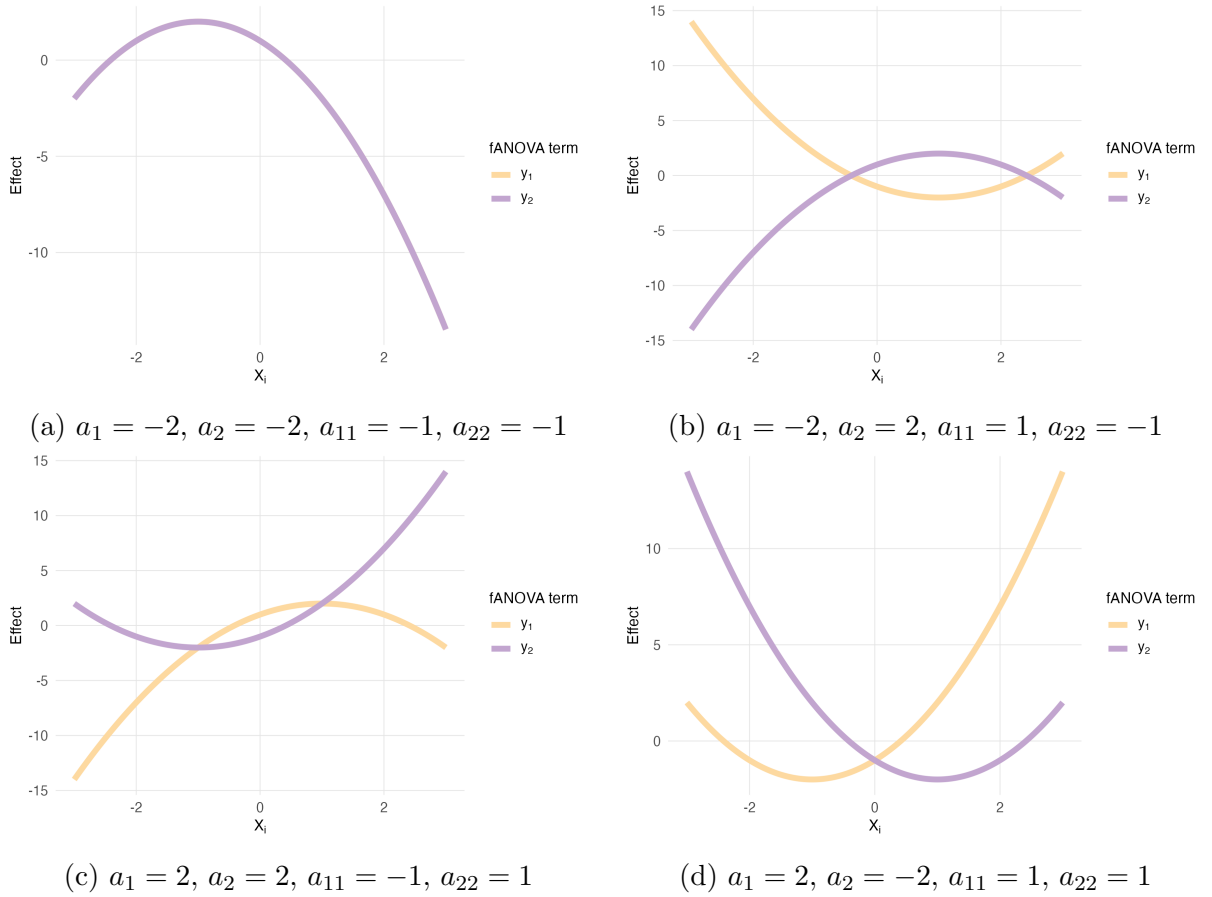

(a) $a_1 = -2$, $a_2 = -2$, $a_{11} = -1$, $a_{22} = -1$

(b) $a_1 = -2$, $a_2 = 2$, $a_{11} = 1$, $a_{22} = -1$

(c) $a_1 = 2$, $a_2 = 2$, $a_{11} = -1$, $a_{22} = 1$

(d) $a_1 = 2$, $a_2 = -2$, $a_{11} = 1$, $a_{22} = 1$

Figure 7: Main effects for different coefficient combinations of mixed main effects. The commponents are given by: $y_1(x_1) = a_1 x_1 + a_{11}(x_1^2 - 1)$, $y_2(x_2) = a_2 x_2 + a_{22}(x_2^2 - 1)$.

Scenario: All

Finally full example, including all main and interaction effects:

$$g_5(x_1, x_2) = a_1 x_1 + a_2 x_2 + a_{11} x_1^2 + a_{22} x_2^2 + a_{12} x_1 x_2.$$

Now the fANOVA components are given by Equation 16, where $a_0 = 0$. We can vary the coefficients as well as ρ.

We recognize the patterns from previous examples which had less parameters to vary. Whether the parabolas are facing upwards or downwards is determined by the sign of the quadratic coefficients. Whether they are facing in the same direction is determined by between variable sign-difference of the quadratic coefficients. Under presence of an interaction term, the direction in which the main effects are facing is also influenced by ρ. The linear coefficients only have an influence on how stretched or compressed the parabolas are. The interaction term is influenced by ρ and a_{12}.

The basis representation is still correct for other distribution assumptions in the sense that it recovers the original function g; however, the component function would not be hierarchically orthogonal.

### 5.2.1 Estimation of fANOVA components

Current landscape of estimation approaches, We need more stable estimation approaches: Especially with real world blackbox models the individual components can be very complicated, probably no analytical closed form themselves The fANOVA decomposition has a strong theoretical foundation but especially in modern work, estimation approaches and computational feasibility is an important aspect to consider. Also, we saw that obtaining the fANOVA components analytically is not always possible, especially under dependent inputs. Therefore, we need to look at estimation approaches that allow us to compute the fANOVA components from data.

**Estimation based on Partial Dependence**

In his estimation framework Hooker (2004) picks up the role of projections in fANOVA. To obtain the component estimate for y_u, he proposes to estimate the projections of y onto the subspace of variables spanned by u empirically. More concretely, one first estimates the conditional expected value of the variables in u (keep variables in u fixed an average over all others). 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 can then be used to estimate the empirical projection of interest. He states that his method works well for functions that have a nearly additive true structure and purely additive functions
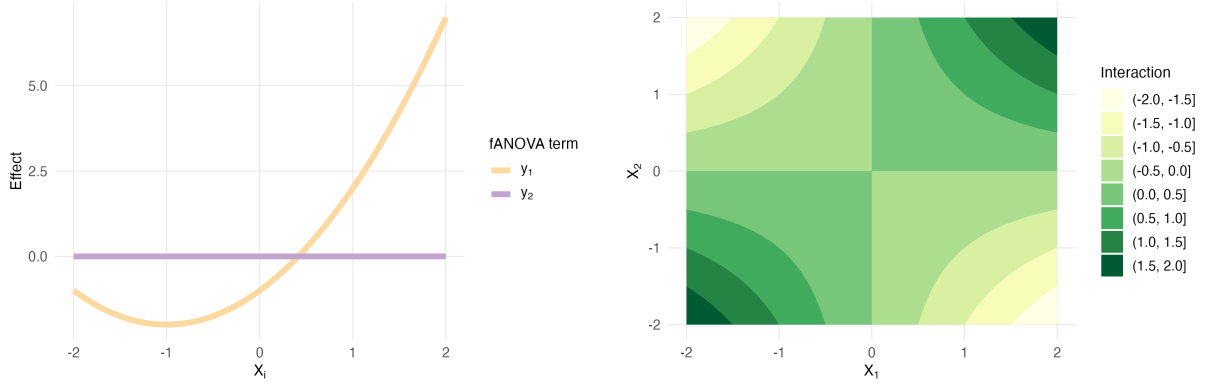
(a) Main and interaction effects for (1) $a_1 = 2$, $a_2 = 0$, $a_{11} = 1$, $a_{22} = 0$, $a_{12} = 0.5$, $\rho = 0$.
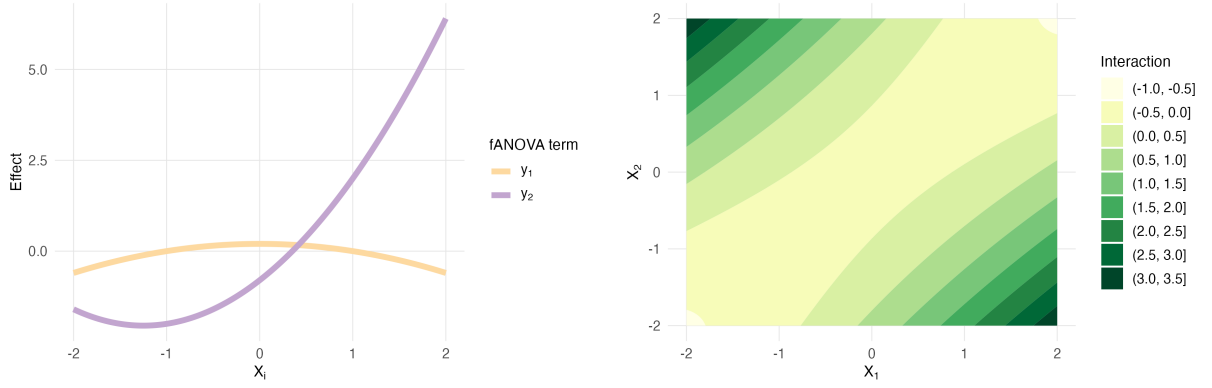
(b) Main and interaction effects for (2) $a_1 = 0$, $a_2 = 2$, $a_{11} = 0$, $a_{22} = 1$, $a_{12} = -0.5$, $\rho = 0.5$.
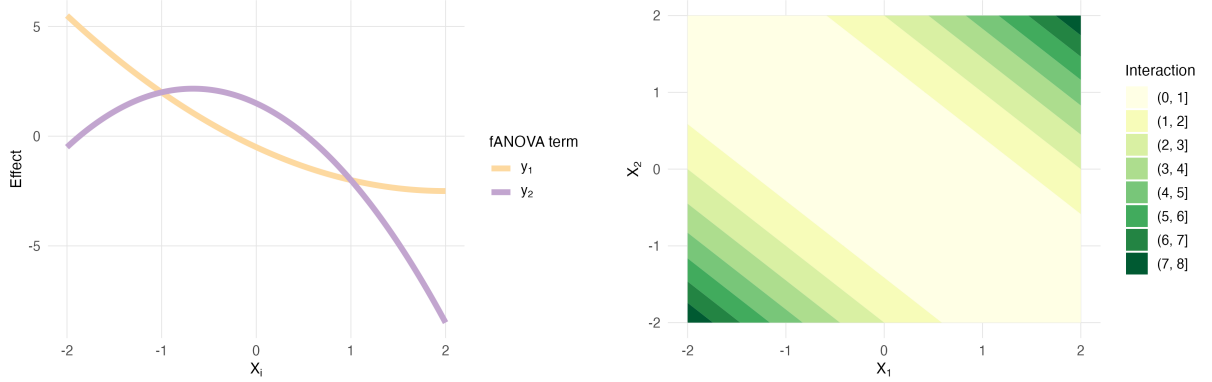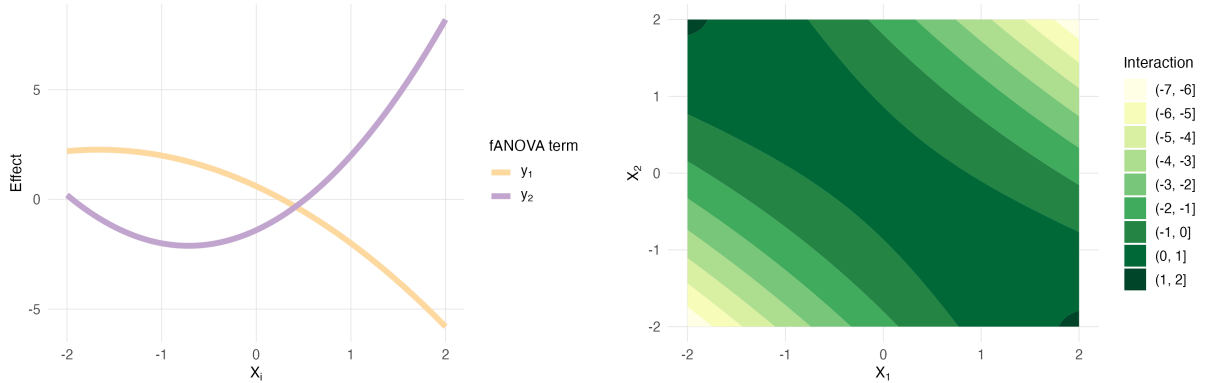

(c) Main and interaction effects for (3) $a_1 = -2$, $a_2 = -2$, $a_{11} = 1$, $a_{22} = -1$, $a_{12} = 1$, $\rho = -1$.

(d) Main and interaction effects for (4) $a_1 = -2$, $a_2 = 2$, $a_{11} = -1$, $a_{22} = 1$, $a_{12} = -1$, $\rho = -0.5$.

Figure 8: Main effects (left) and interaction contours (right) for four different coefficient sets.

are exactly recoverable with this approach. To save computational costs, he proposes to base the Monte Carlo estimates of the PD function on a randomly sampled subset of data points. Problem: no true projections (under dependence or always?); extrapolation issues etc.; even if no product type measure assumption, still problems in handling dependent inputs.

Estimation based on weighted least squares

Hooker (2007) proposes a new estimation scheme for his generalized fANOVA decomposition. The mathematical problem one faces is more complex: the fANOVA components are defined in dependence of each other and system has to be solved simultaneously as we saw in the previous section. Hooker rewrites the estimation problem as a restricted weighted least squares problem and solves it via Lagrange multiplier for the exact solution of the simultaneously defined generalized components; problem restricted to ensure hierarchical orthogonality. The function is again evaluated at a grid of points to reduce the problem to a finite dimensional one. Because of the parallel to weighted least squares, it is also possible to compute a weighted standard ANOVA with existing software, but it is difficult to incorporate the constraints, so the components may not be hierarchical orthogonal.

None of these estimation approaches has a standard software implementation or published code. Some existing more or less finished implementations are numerically instable or yield illogical results. For our following experimental setup we therefore move within the range bivariate polynomials of degree two offer us and for which the solution is provided by Rahman.

6 Conclusion

We started by working through the historical context of the fANOVA decomposition. We explored the origins of the fANOVA method rooted in mathematical work by Hoeffding (1948) and Sobol (1993). We saw how the method was picked up by following researchers in different contexts.

Clear contribution of this work: brought clarity and unity to the various different formulations of fANOVA. We see trend in recent ML literature (cite all these ML papers with the fancy models), pick up the methods but the theoretical background and clean formalism often left aside. This work serves as a reference to practitioners who seek a unified and clean formalization of the fANOVA method. Filled the void of visualizations and intuitions around the method due to the lack of software implementations.

Outlook, work that could follow from this thesis: Examine the different approaches to estimate fANOVA components (how do they scale? what is their accuracy? etc.) Write software implementation for fANOVA decomposition; current landscape is sparse but the method has great potential for IML; with current practicability it is however clear that fANOVA will not be accepted, it is not convenient to use the method fANOVA powerful theory, sound mathematical foundation, but without standardized software implementation application to IML difficult. Parallels to Shapley values, unified under a game theoretic approach; Fumagalli et al. (2025) recently established this parallel, would be very interesting to investigate further.

A   Appendix

Proof of classical fANOVA decomposition

Here we show the 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:*

$$y(\boldsymbol{X}) = \sum_{u \subseteq \{1,\ldots,N\}} y_u(\boldsymbol{X}_u), \tag{17}$$

subject to the zero mean constraint Equation 3.

Sobol proofs existence and uniqueness of the fANOVA decomposition by showing how the summands of the desired decomposition look and showing that they satisfy the desired zero mean property.

Proof. Assume that $\boldsymbol{X}$ is an N-dimensional vector of independent random variables and that the zero mean property holds for the - still abstract - fANOVA terms. We define the integral w.r.t. all variables except for the ones with indices in v:

$$g_v(\boldsymbol{x}_v) = \int_{[0,1]^{N-|v|}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}_{-v}) \tag{18}$$

The very first term in the decomposition is the integral of y with respect to all variables:

$$y_\emptyset = \int_{\mathbb{R}^N} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) \tag{19}$$

This integral exists because $y \in L^2(\mathbb{R}^N, f_{\boldsymbol{X}} d\nu)$, and the product measure is finite on the domain.

Next, Sobol derives the one dimensional fANOVA terms. For this, take the integral of

Equation 1 w.r.t. all variables except for the one with index i, so $v_1 = \{i\}$:

$$\int_{\mathbb{R}^{N-1}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) = \int_{\mathbb{R}^{N-1}} \sum_{u \subseteq \{1,\dots,N\}} y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) \tag{20}$$

$$= \sum_{u \subseteq \{1,\dots,N\}} \int_{\mathbb{R}^{N-1}} y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) \tag{21}$$

$$= \sum_{u \subseteq \{1,\dots,N\}} \int_{\mathbb{R}^{N-1}} y_u(\boldsymbol{x}_u) \left(\prod_{j=1}^{N} f_{X_j}(x_j) \right) d\nu(\boldsymbol{x}_{-v_1}) \tag{22}$$

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:

$$\int_{\mathbb{R}^{N-1}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) = y_\emptyset + y_{\{i\}}(x_i). \tag{23}$$

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

$$y_{\{i\}}(x_i) = \int_{\mathbb{R}^{N-1}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v_1}) - y_\emptyset. \tag{24}$$

Now consider $v_2 = \{1,2\}$. We integrate the ANOVA decomposition over all variables except x_1 and x_2:

$$\int_{\mathbb{R}^{N-2}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-\{1,2\}}) = \int_{\mathbb{R}^{N-2}} \sum_{u \subseteq \{1,\dots,N\}} y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-\{1,2\}}) \tag{25}$$

$$= \sum_{u \subseteq \{1,\dots,N\}} \int_{\mathbb{R}^{N-2}} y_u(\boldsymbol{x}_u) \left(\prod_{j=1}^{N} f_{X_j}(x_j) \right) d\nu(\boldsymbol{x}_{-\{1,2\}}) \tag{26}$$

$$= y_\emptyset + y_{\{1\}}(x_1) + y_{\{2\}}(x_2) + y_{\{1,2\}}(x_1,x_2) \tag{27}$$

Hence, the two-dimensional component is given by:

$$y_{\{1,2\}}(x_1,x_2) = \int_{\mathbb{R}^{N-2}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-\{1,2\}}) - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2) \tag{28}$$

We can continue this process for all combinations of indices $v \subseteq \{1, \ldots, N\}$ 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:

$$y_v(\boldsymbol{x}_v) = \int_{\mathbb{R}^{N-|v|}} y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}_{-v}) - \sum_{u \subsetneq v} y_u(\boldsymbol{x}_u) \tag{29}$$

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

$$y_{\{1,\ldots,N\}}(\boldsymbol{x}) = y(\boldsymbol{x}) - \sum_{u \subsetneq \{1,\ldots,N\}} y_u(\boldsymbol{x}_u) \tag{30}$$

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

$$\int y_v(\boldsymbol{x}_v) f_{X_i}(x_i)\, d\nu(x_i) = \int \left(\int y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-v}) - \sum_{u \subsetneq v} y_u(\boldsymbol{x}_u) \right) f_{X_i}(x_i)\, d\nu(x_i) \tag{31}$$

$$= \int \left(\int y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-v}) \right) f_{X_i}(x_i)\, d\nu(x_i) - \sum_{u \subsetneq v} \int y_u(\boldsymbol{x}_u) f_{X_i}(x_i)\, d\nu(x_i) \tag{32}$$

$$= \int y(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\nu(\boldsymbol{x}_{-v}) d\nu(x_i) - \sum_{u \subsetneq v} \int y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u)\, d\nu(\boldsymbol{x}_u) \tag{33}$$

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:

$$\int y_v(\boldsymbol{x}_v) f_{X_i}(x_i)\, d\nu(x_i) = y_\emptyset - y_\emptyset = 0 \tag{34}$$

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

$$\int y_v(\boldsymbol{x}_v) f_{X_i}(x_i)\, d\nu(x_i) = 0, \quad \text{for all } i \in v. \tag{35}$$

This completes the proof of the existence and uniqueness of the fANOVA decomposition,

as well as the verification of the zero mean property for all components.   □

Square Integrability of $f_1(x_1)$

For now we want to show that the single fANOVA term $f_1(x_1)$ is square integrable, given that the original function $f(x) \in \mathcal{L}^2$. We need to show that:

$$\int |f_1(x_1)|^2 \, dx_1 < \infty$$

The single fANOVA term is defined as:

$$f_1(x_1) = \int f(x) \, dx_{-1} - f_0$$

We take the squared norm, and integrate w.r.t. x_1 to use the Cauchy-Schwarz inequality:

$$\int |f_1(x_1)|^2 \, dx_1 = \int \left| \int f(x) \, dx_{-1} - f_0 \right|^2 dx_1$$

$$= \int |(\int f(x) \, dx_{-1})^2 - 2 \int f(x) \, dx_{-1} f_0 + f_0^2| dx_1$$

Break this into three terms:

$$(1): \quad \int \left| \int f(x) \, dx_{-1} \right|^2 dx_1 \leq \int \left(\int 1^2 \, dx_{-1} \right) \left(\int |f(x)|^2 \, dx_{-1} \right) dx_1 = \int |f(x)|^2 \, dx < \infty$$

$$(2): \quad 2 \int \left(\int f(x) \, dx_{-1} \right) f_0 \, dx_1 = 2 f_0 \int \left(\int f(x) \, dx_{-1} \right) dx_1 = 2 f_0^2 < \infty$$

$$(3): \quad \int f_0^2 \, dx_1 = f_0^2 < \infty$$

Since each term (1)–(3) is finite, and $\int |f_1(x_1)|^2 dx_1$ is a linear combination of them: $\int |f_1(x_1)|^2 dx_1 < \infty$

B Electronic appendix

Data, code and figures are provided in electronic form.

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**(none).

Choi, Y., Park, S., Park, C., Kim, D. and Kim, Y. (2025). Meta-anova: screening interactions for interpretable machine learning, *Journal of the 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 [cs].
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. (2024). 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.

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.

Location, date

Name