Bachelor's Thesis

---

fANOVA for Interpretable Machine Learning

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, Month Day$^{\text{th}}$, Year

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

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Contents

1 Introduction

Problem: fANOVA decomposition such a foundational concept, long history, mix of formalizations; This starts with the name of the method (decomposition into summands of different order, ANOVA representation, functional ANOVA (fANOVA) decomposition Hooker (2004), ANOVA dimensional decomposition (ADD)) And continues with how fANOVA is formalized in the literature; each paper uses different notation, deviating settings, set of assumptions, some use formulation via expected value, other via integral,. All of this equivalent and comes together under concept of orthogonal projections, but this is hard to see when one start learning about fANOVA. At the core fANOVA provides a way in which integrable functions can be decomposed into a sum of orthogonal components. Because of tits foundational nature fANOVA is useful in interpretability of machine learning models (Hooker, Molnar), uncertainty quantification of complex systems (Rahman (2014)), non-parametric statistical modelling (stone etc.), sensitivity analysis (Sobold), and many more fields. This also means that there exists a need for 1. A comprehensive overview of fANOVA-related work and 2. To clean up various notations and definitions that in the end state the same. Brining clarity into the fANOVA landscape is more relevant than ever as fANOVA decomposition has attracted attention in recent IML literature; used to build fancy models, but the theoretical foundation often goes unaddressed. Which is understandable when the other end of the spectrum is formed by rigorous, but often inaccessible, theoretical papers, heavy from measure theoretic viewpoint. In between we have work that is mathematically clean in itself but makes underlying assumption that come not across resulting in pracitioners that do not understand what assumptions they are making when using the expression.

Possible Solution: This paper aims to provide 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 practitioners to understand the method, its mathematical background, and also provide an overview for how the method is used in other research context. This work is organized as follows: It starts with historical context and how the method has evolved over time; we then give formal introduction to classical fANOVA as well as the generalization to dependent inputs. We will outline possible estimation schemes, particularly relevant for an application of the method. Next we illustrate characteristics of the classical fANOVA based on analytical examples; before concluding with a discussion of the method's limitations and possible future research directions.

2   Background Knowledge

Basic Setup

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. $\boldsymbol{X_u} = (X_1, \ldots, X_{|u|}), u \neq \emptyset, 1 \leq i_1 < \cdots < i_{|u|} \leq N$ is a subvector 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 $\emptyset \neq u \subseteq \{1, \ldots N\}$. $f(\boldsymbol{X}) := f(X_1, \ldots, X_N)$ is a mathematical model with random variables as inputs. We write a vector space of square-integrable functions as

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

.

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

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

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

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

We denote strict inclusion by $\subsetneq$ and $\subset$ allows for equality. Formal setup now based on Chastaing et al. (2012), Rahman (2014). Maybe also cite: https://apachepersonal.miun.se/ andrli/Bok.pdf; also look again at pdf and measure setup.

Orthogonal projection

Let $\mathcal{G} \subset \mathcal{L}^2$ denote a linear subspace. The projection of f onto $\mathcal{G}$ is defined by the function $\Pi_{\mathcal{G}} f$ which minimizes the distance to f in $\mathcal{L}^2$:

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

Definition of $\mathcal{L}^2$ space and projection modified from https://tnagler.github.io/mathstat-lmu-2024.pdf.

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:

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

Only when X_1 and X_2 are independent can we write

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

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

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

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 History of fANOVA

3.1 Early Work on fANOVA

The main idea of the fANOVA decomposition is to decompose a statistical model into the sum of the main effects and interaction effects of its input variables. The underlying principle of fANOVA decomposition dates back to Hoeffding (1948). In his seminal work(?) on estimators with asymptotical normal distribution, he introduced U-statistics, along with the "Hoeffding decomposition", which allows to write a symmetric function of the data as a sum of orthogonal components. Sobol (1993) used the same principle and applied it to deterministic mathematical models. proofed existance of fANOVA decomposition for square integrable functions. He built on the originally called "decomposition into summands of different dimension" in Sobol (2001), where he introduces Sobol indices and renames the method to the "ANOVA-representation". For Sobol decomposing the function into the sum of fANOVA terms is actually not central, but what he is mostly interested is the variance decomposition which he shows follows from the fANOVA decomposition of a function. This variance decomposition allows quantifying how much the variance of a single input variable contributes to the overall variance of the function. Thus, Sobol indices are commonly used in sensitivity analysis. Sobol builds his main contributions around fANOVA on the 1) variance decomposition, but also proposes to use fANOVA for 2) variable selection/ dimensionality reduction (terms that contribute a lot to overall variance should be in the model).

Efron and Stein (1981) use the idea of the decomposition to proof their famous lemma on jackknife variances.

A true wave of fANOVA literature around the 1990s, where authors investigate fANOVA-based models, establish parallels to splines, study their theoretical properties (convergence, consistency, etc.), and prectical use cases (dimensionality reduction, etc.). All cited in Huang (1998b). Stone (1994) mainly uses fANOVA decomposition to base smooth regression models with interactions on it and his paper is the building block for a broader body of work of fANOVA-based models (see for example Huang (1998a,b)) Go deeper into some of the works? And a wrap up? Wrap up: fANOVA received a lot of attention in statistics literature, its mathematical properties were studied and was also used as intermediate step to proof or build other theories.

3.2 Modern Work on fANOVA

The fANOVA decomposition has a long history with roots in mathematical statistics and non-parametric estimation theory.

Owen (2013) formal intro to fANOVA decomposition and generalization of Sobol indices. Owen has generally a lot of work related to fANOVA decomposition, either lecture notes explaining the decomposition, methods based on it Owen (2003), or deeper into sensitivity analysis and fANOVA Owen (2013).

Since the assumptions of independent variables in classical fANOVA is often too restrictive in practice, Hooker (2007) generalizes the method to dependent variables. A recent paper by Il Idrissi et al. (2025) can be seen as another approach to generalize the principle of fANOVA decomposition to dependent inputs.

In more recent years, the method has been rediscovered by the machine-learning community, especially in the context of interpretable machine learning (IML) and explainable AI (XAI). Hooker (2004) introduces the fANOVA decomposition with the goal of providing a global explanation method for black-box models. And recent work discovered interesting mathematical parallels between fANOVA and other IML methods, such as PDP Friedman (2001), or Shapley values (Fumagalli et al. (2025), Herren, Owen preprint).

GA2ML etc.

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). Liu and Owen (2006) use of fANOVA and sensitivity analysis for functions arising in computational finance.

fANOVA/ variance decomposition to reduce dimensionality, fANOVA to find additive structures (explanability and surrogate modelling), fANOVA for identifying interaction terms and variable dependencies.

4 Classical fANOVA

4.1 Formal Introduction to fANOVA

(Sobol (1993) proofs the existence of the classical fANOVA decomposition; other papers give definitions we base our definition on.)This chapter is based on the formal introductions by Rahman (2014), Sobol (1993, 2001), Hooker (2004), Owen (2013), Muehlenstaedt et al. (2012). We show both formulations of the fANOVA, via the integral and via the expected value and in general prefer the expected value formulation as it is more intuitive in a probabilistic setting. Originally, Sobol (1993) presented the fANOVA decomposition with independent input variables with support bounded to the unit interval, i.e. he considered the measure space $([0,1]^n, \mathcal{B}([0,1]^n), \nu)$. Later work shows that this restriction is not necessary, and we can work with the Borel σ-algebra on the n-dimensional real number line, i.e. $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \nu)$, and with a general measure ν defined on it (see e.g. Rahman (2014)). Since we assume independence of the input variables, their joint distribution is given by the product over the marginal distributions, i.e. $f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) = \prod_{i=1}^{N} f_{X_i}(x_i) \, d\nu(x_i)$. $f_{X_i} : \mathbb{R} \to \mathbb{R}_0^+$ is the marginal probability density function of X_i defined on $(\Omega_i, \mathcal{F}_i, \nu_i)$.

Definition 4.1. *Let y denote a mathematical model with realizations of independent random variables $x_1, \ldots, x_N$ as input. We can represent such a model y as the hierarchical sum of specific basis functions with increasing dimensionality:*

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

If $|u| = 0$ it describes the constant term, if $|u| = 1$ it describes the main effects, if $|u| > 1$ it describes the interaction effects of the variables in u. The expansion consists of 2^N terms.

Construction of the fANOVA Terms

The individual fANOVA term for the variables with indices in u are constructed from 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 accounting 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:

$$y_\emptyset = \int y(\boldsymbol{x}) \prod_{i=1}^{N} f_{X_i}(x_i) \, d\nu(x_i) = \mathbb{E}[y(\boldsymbol{X})]. \tag{2}$$

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

$$y_u(\boldsymbol{X}_u) = \int y(\boldsymbol{X}_u, \boldsymbol{x}_{-u}) \prod_{i=1, i \notin u}^{N} f_{X_i}(x_i) \, d\nu(x_i) - \sum_{v \subsetneq u} y_v(\boldsymbol{X}_v), \tag{3}$$

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

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 explanation method of black box models.

The fANOVA terms should be constructed in such a way that they have two specific properties crucial for identifiability and interpretation.

Proposition 4.1. *The zero-mean property states that all effects, except for the constant terms, are centred around zero. Mathematically this means that the effects integrate to zero w.r.t. their own variables:*

$$\int y_u(\boldsymbol{x}_u) f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) := \mathbb{E}[y_u(\boldsymbol{X}_u)] = 0 \tag{4}$$

Proposition 4.2. *The second property is the orthogonality of the fANOVA terms. 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 their fANOVA terms are orthogonal to each other:*

$$\int y_u(\boldsymbol{x_u}) y_v(\boldsymbol{x_v}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \mathbb{E}[y_u(\boldsymbol{X}_u) y_v(\boldsymbol{X}_v)] = 0 \tag{5}$$

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. Rahman (2014) derives these two properties (Equation 4, Equation 5)from a more general condition, he calls the "strong annihilating conditions".

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

probability density functions:

$$\int y_u(\boldsymbol{x}_u) f_{X_i}(x_i) \, d\nu(x_i) = 0, \quad \text{for } i \in u \neq \emptyset. \tag{6}$$

We can reassure ourselves that the properties in fact follow from the strong annihilating conditions. For the zero-mean constraint we can write:

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

One can follow the same reasoning for the orthogonality condition:

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

4.2 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:

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

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 constraint, 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:

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

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.

4.3 fANOVA as projection

In the following we revisit the fANOVA decomposition from the view of orthogonal projections. The section is based on Vaart (1998). This will also help to understand the generalization of fANOVA in section 5.

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 (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 (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 $u \subseteq \{1, \ldots, N\}$ 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

We already established that $\mathbb{E}[y(\boldsymbol{X})] = y_\emptyset$. For the variance of $y(\boldsymbol{X})$, we find that the total variance can be decomposed into the sum of the fANOVA term variances. The variance

decomposition is a major result in Sobol (1993) and forms the basis for the Sobol indices in sensitivity analysis. We sketch the variance decomposition here and note that it is only possible under independence assumption.

If $y \in \mathcal{L}^2$, then $y_{i_1, \dots, y_n} \in \mathcal{L}^2$ proof? reference?; Sobol 1993 says it is easy to show using Schwarz inequality and the definition of the single fANOVA terms. Therefore, we define the variance of f as follows:

$$\sigma^2 := \int y^2(\boldsymbol{X}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}) - (y_0)^2$$
$$= \int y^2(\boldsymbol{X}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}) - (\int y(\boldsymbol{X}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(\boldsymbol{x}))^2$$
$$= \mathbb{E}[y^2(\boldsymbol{X})] - \mathbb{E}[y(\boldsymbol{X})]^2$$

The variance of the fANOVA components is then defined as

$$\sigma^2_{x_{i_1}, \dots, x_{i_n}} = \int \cdots \int y^2_{i_1, \dots, i_n} f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(x_1) \cdots d\nu(x_n)$$
$$- \left(\int \cdots \int f_{i_1, \dots, i_n} f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(x_1) \cdots d\nu(x_n) \right)^2$$
$$= \mathbb{E}[y^2_{i_1, \dots, i_n}] - \mathbb{E}[y_{i_1, \dots, i_n}]^2$$

Because of the orthogonality property, the second term vanished and we get:

$$\sigma^2_{x_{i_1}, \dots, x_{i_n})} = \int \cdots \int y^2_{i_1, \dots, i_n} f_{\boldsymbol{X}}(\boldsymbol{x}) d\nu(x_1) \cdots d\nu(x_n)$$
$$= \mathbb{E}[y^2_{i_1, \dots, i_n}]$$

With the definition of the total variance σ^2 and the component-wise variance $\sigma^2_{x_{i_1}, \dots, x_{i_n}}$ we can now see that the total variance can be decomposed into the sum of the component-wise variances.

Alternatively we can formulate this via the expected value. We write the sum over u for the sum over $\emptyset \neq u \subseteq \{1, \dots, N\}$ and the sum over $u \neq v$ for the sum over $\emptyset \neq u \subseteq$

$\{1, \ldots, N\}, \emptyset \neq v \subseteq \{1, \ldots, N\}, u \neq v.$

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

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

$$Var(a + X_1 + 2X_2 + X_1X_2) = Var(X_1) + 4Var(X_2) + Var(X_1X_2) + 2Cov(X_1, X_2)$$
$$= 1 + 4 \cdot 1 + 1 \cdot 1 + 2 \cdot 0 = 6$$
$$= \mathbb{E}[X_1^2] + 4\mathbb{E}[X_2^2] + \mathbb{E}[X_1^2]\mathbb{E}[X_2^2] + 2Cov(X_1, X_2)$$
$$= \mathbb{E}[y_1^2] + \mathbb{E}[y_2^2] + \mathbb{E}[y_{12}^2]$$

5 Generalized fANOVA

5.1 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 4.3. 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:

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

The fANOVA components are characterized by two central properties zero mean and orthogonality which follow from Equation 6. 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:

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

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 the whole point of the fANOVA decomposition and central so that it adds sth. IML.

5.2 Formal Introduction to 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 $(X_1, \ldots, X_n)$.

Definition 5.1. *Generalized fANOVA decomposition.* *We denote the generalized functional fANOVA decomposition as:*

$$y(\boldsymbol{X}) = \sum_{u \subseteq \{1,\ldots,N\}} y_{u,G}(\boldsymbol{X}_u) \tag{7}$$

The subscript G indicates that we are working with the generalized fANOVA components. The main question is, how one can build fANOVA components that still satisfy the desired

properties of zero mean and orthogonality under dependent inputs.

To ensure that these statements hold for the generalized fANOVA components, we need to set the weak annihilating conditions. They fulfill the same function as the strong annihilating conditions do in the classical case but work with the joint density of the variables of interest, instead of the individual marginal probability density functions. This makes sense, because when there are dependencies between variables then the individual pdfs would not assign the correct weight to each function value as they ignore the dependence between features in u.

Proposition 5.1. _Weak annihilating conditions._ _To ensure the two desired properties of the generalized fANOVA components (zero mean, hierarchical orthogonality), we require the weak annihilating conditions:_

$$\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u) d\nu(x_i) = 0 \quad for \quad i \in u \neq \emptyset \tag{8}$$

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.

Proposition 5.2. _The generalized fANOVA components $y_{u,G}$, with $\emptyset \neq u \subseteq \{1, \ldots, N\}$, are centred around zero:_

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

Proof. For any subset $\emptyset \neq u \subseteq \{1, \ldots, N\}$, 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 $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

Proposition 5.3. *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:*

$$\mathbb{E}[y_{u,G}(\boldsymbol{X}_u)y_{v,G}(\boldsymbol{X}_v)] := \int y_{u,G}(\boldsymbol{x}_u)y_{v,G}(\boldsymbol{X}_v)f_{\boldsymbol{X}}(\boldsymbol{x})\,d\nu(\boldsymbol{x}) = 0 \qquad (10)$$

Proof. 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

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

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

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. We can show that 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. $\qquad \square$

Construction of the Generalized fANOVA Terms

Rahman basically says, that there is no need for a computationally expensive least squares problem proposed by Hooker (2007). In the same fashion as for the classical fANOVA decomposition, one can integrate the desired decomposition w.r.t. to a suitable pdf and use the (weak) annihilating conditions to obtain the fANOVA decomposition. They key is to find this suitable pdf which results in an integral as we want it. Rahman (2014) says this suitable pdf is the $f_{-u}(\boldsymbol{x}_{-u})$. In Theorem 4.4. he shows how the generalized fANOVA components constructed by this look like. And Lemma 4.3 is a helping statement, that should cover all the integral cases that appear in Theorem 4.4. and allow us to write the expressions in Theorem 4.4. in the reduced form we see them. So I should be able to find the cases of Lemma 4.3. in the building of Theorem 4.4.

To construct the generalized fANOVA components we first need a lemma that dinstiguishes three cases of integration that we will encounter in the construction of the generalized components.

Lemma 5.1. *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.

$\square$

Theorem 5.1. *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.

$\square$

Hooker (2007) offers an alternative definition of the generalized fANOVA components[1]:

$$\{y_{u,G}(\boldsymbol{x}_u) \mid u \subseteq d\} = \arg\min_{\{g_u \in L^2(\mathbb{R}^u)\}_{u \subseteq d}} \int \left(\sum_{u \subseteq d} g_u(\boldsymbol{x}_u) - y(\boldsymbol{x}) \right)^2 f_{\boldsymbol{X}}(\boldsymbol{x}) \, d\nu(\boldsymbol{x}) \qquad (11)$$

In Hookers definition we recognize a projection problem. 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 2.

A crucial difference to the classical case is that both versions of the generalized components are defined in dependence of each other (**??**, Equation 11).

Let us come back to our example from the beginning. The goal is to write

$$g(x_1, x_2) = y_{\emptyset,G} + y_{1,G}(x_1) + y_{2,G}(x_2) + y_{1,2,G}(x_1, x_2)$$

under dependent inputs. We present two ways in which the problem solution can be stated.

---

[1]We modified the notation from the original work to match the notation of Rahman (2014) and the rest of this thesis.

Rahman method

The system to find the generalized fANOVA components for g according to Rahman (2014) method looks as follows:

$$y_{\emptyset,G} = \int_{\mathbb{R}^2} g(x_1, x_2)\, f(x_1, x_2)\, dx_1 dx_2$$

$$y_{1,G}(x_1) = \int_{\mathbb{R}} g(x_1, x_2)\, f_2(x_2)\, dx_2 - y_{\emptyset,G} - \int_{\mathbb{R}} y_{\{1,2\},G}(x_1, x_2)\, f_2(x_2)\, dx_2$$

$$y_{2,G}(x_2) = \int_{\mathbb{R}} g(x_1, x_2)\, f_1(x_1)\, dx_1 - y_{\emptyset,G} - \int_{\mathbb{R}} y_{\{1,2\},G}(x_1, x_2)\, f_1(x_1)\, dx_1$$

$$y_{1,2,G}(x_1, x_2) = g(x_1, x_2) - y_{\emptyset,G} - y_{\{1\},G}(x_1) - y_{\{2\},G}(x_2)$$

Since the components form a coupled system where the components are defined in interdependence of each other, finding the solution is not straight forward, even for simple examples.

Hooker method

An alternative way to phrase the problem can be found in Hooker (2007). To find the generalized fANOVA components, we can formulate a minimization problem for each of them.

$$y_{\emptyset} = \arg\min_{c \in \mathbb{R}} \int_{\mathbb{R}^2} \left(g(x_1, x_2) - \left(c + y_{\{1\}}(x_1) + y_{\{2\}}(x_2) + y_{\{1,2\}}(x_1, x_2) \right) \right)^2 f(x_1, x_2)\, dx_1 dx_2$$

$$y_1(x_1) = \arg\min_{h_1 \in L^2(\mathbb{R})} \int_{\mathbb{R}^2} \left(g(x_1, x_2) - \left(y_{\emptyset} + h_1(x_1) + y_{\{2\}}(x_2) + y_{\{1,2\}}(x_1, x_2) \right) \right)^2 f(x_1, x_2)\, dx_1 dx_2$$

$$y_2(x_2) = \arg\min_{h_2 \in L^2(\mathbb{R})} \int_{\mathbb{R}^2} \left(g(x_1, x_2) - \left(y_{\emptyset} + y_{\{1\}}(x_1) + h_2(x_2) + y_{\{1,2\}}(x_1, x_2) \right) \right)^2 f(x_1, x_2)\, dx_1 dx_2$$

$$y_{1,2}(x_1, x_2) = \arg\min_{h_{12} \in L^2(\mathbb{R}^2)} \int_{\mathbb{R}^2} \left(g(x_1, x_2) - \left(y_{\emptyset} + y_{\{1\}}(x_1) + y_{\{2\}}(x_2) + h_{12}(x_1, x_2) \right) \right)^2 f(x_1, x_2)\, dx_1 dx_2$$

The least-squares problems are solved subject to the following constraints, which ensure

that the resulting components are zero centred and hierarchically orthogonal:

$$\int_{\mathbb{R}^2} y_{\{1\}}(x_1) \cdot f(x_1, x_2)\, dx_1 dx_2 = 0$$

$$\int_{\mathbb{R}^2} y_{\{2\}}(x_2) \cdot f(x_1, x_2)\, dx_1 dx_2 = 0$$

$$\int_{\mathbb{R}} y_{\{1,2\}}(x_1, x_2) \cdot f(x_1, x_2)\, dx_1 = 0 \quad \forall x_2$$

$$\int_{\mathbb{R}} y_{\{1,2\}}(x_1, x_2) \cdot f(x_1, x_2)\, dx_2 = 0 \quad \forall x_1$$

Conceptually Hooker (2007) is doing nothing other than a projection. Earlier, we established that a projection is the same as the conditional expected value, and fANOVA can be expressed via the conditional expected value. This means from the initial idea, we do not change anything apart from the fact that we have to integrate via the joint pdf, but this is something one is "forced" to under dependence, not something one "invents". However, projections onto subspaces become more difficult under dependence; therefore, setting these constraints explicitly is necessary to ensure (hierarchical) orthogonality.

Obtaining an analytical solution for either of the methods is tedious even for our simple example. We leave it at the problem formulation, so that we have the comparison between which problem one has to solve the classical case versus the generalized case. In the next section, we sketch ways to estimate the fANOVA components conceptually.

Second-moment statistics

As we saw earlier, the expected value of the generalizes fANOVA decomposition is the constant term:

$$\mathbb{E}[y(\mathbf{X})] = y_{\emptyset, G}. \tag{12}$$

The variance of the generalized fANOVA decomposition can be expressed as follows:

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

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.

6 Estimation of fANOVA

The fANOVA decomposition has a strong theoretical foundation but especially in modern work, estimation approaches and computational feasibility is an important aspect to consider.

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 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.

Estimation based on polynomial expansion

Rahman (2014) approaches the estimation differently and represents each component with a Fourier polynomial expansion. For normally distributed inputs, one can choose Hermite polynomials as basis functions, which simplifies things; in other cases more complicated I think. Again need to look up the mechanisms better. Other work, quiet mathematical that also goes in this direction I think.

No standard implementations available. Use existing software for parts (e.g. Monte Carlo estimates) and workarounds.

7 Examples & Visualizations

Known function, simulate input, plot fANOVA and PD, comparison
PD and fANOVA: If inputs are independent, I think for the main effects fANOVA is simply PD shifted but fANOVA answers a different question thatn PD. PD usually asks for the effect of one specific variable on the prediction. In contrast, fANOVA decomposes the entire function. It is more about a global representation and clean, isolated effects because in sum they have to recover the original function and may not overlap.

Standard MVN, linear function, interaction

The plots show the classical fANOVA components for our known example from the previous sections. Classical components visualized, but in this simple example they are simply the subfunctions we can already see from looking at the function.
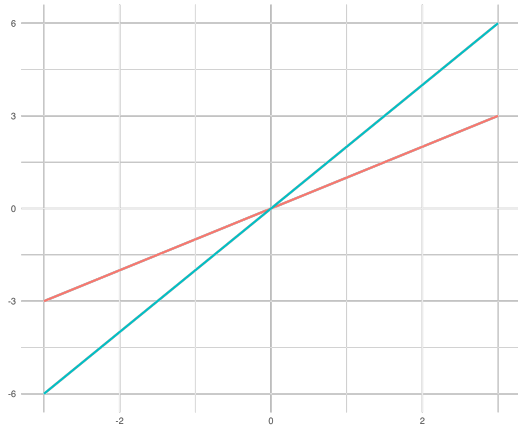

Figure 1: Main terms as calculated via classical fANOVA for $g(x) = x_1 + 2x_2 + x_1 x_2$.
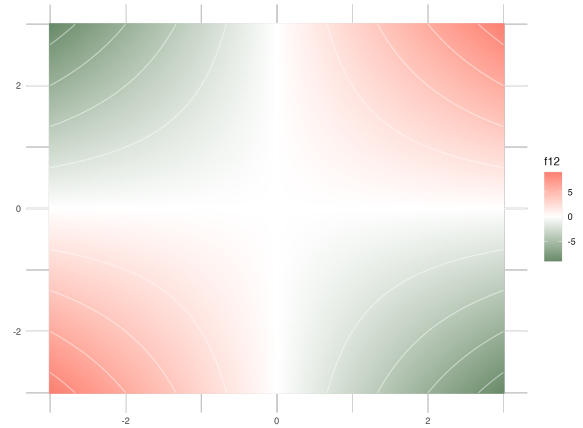
Figure 2: Contour plot of $g(x) = x_1 + 2x_2 + x_1 x_2$.

Next, instead of a standard MVN distribution assumption for the inputs, we allow for non-centred inputs. This is to confirm that the fANOVA decomposition manages to yield zero mean components, even when inputs are not centred. $g = a + X_1 + 2X_2 + X_1 X_2$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right).$$

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:

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}(\mu_1, 1), \quad X_2 \mid X_1 = x_1 \sim \mathcal{N}(\mu_2, 1)$$

We can now compute the classical fANOVA components as follows:

$$
\begin{aligned}
y_\emptyset &= \mathbb{E}[g(X)] = a + \mu_1 + 2\mu_2 + \mu_1\mu_2, \\
y_1 &= \mathbb{E}[g(X) \mid X_2 = x_2] - y_\emptyset = a + 2\mu_2 + x_1 + x_1\mu_2 - y_\emptyset \\
&= x_1(1 + \mu_2) - \mu_1\mu_2 - \mu_1, \\
y_2 &= \mathbb{E}[g(X) \mid X_1 = x_1] - y_\emptyset = a + \mu_1 + 2x_2 + x_2\mu_1 - y_\emptyset \\
&= x_2(2 + \mu_1) - \mu_1\mu_2 - 2\mu_2, \\
y_{12} &= g(x_1, x_2) - y_\emptyset = x_1 x_2 - \mu_2 x_1 - \mu_1 x_2 + \mu_1\mu_2.
\end{aligned}
$$

We recognize that each fANOVA components is shifted by constants (that are formed from the conditional and unconditional expected values of the input variables).
It is easy to verify that non-constant terms have mean zero:

$$
\begin{aligned}
\mathbb{E}[y_1] &= \mathbb{E}[X_1(1 + \mu_2) - \mu_1\mu_2 - \mu_1] = (1 + \mu_2)\mathbb{E}[X_1] - \mu_1\mu_2 - \mu_1 = 0, \\
\mathbb{E}[y_2] &= \mathbb{E}[X_2(2 + \mu_1) - \mu_1\mu_2 - 2\mu_2] = (2 + \mu_1)\mathbb{E}[X_2] - \mu_1\mu_2 - 2\mu_2 = 0, \\
\mathbb{E}[y_{12}] &= \mathbb{E}[X_1 X_2] - \mu_2\mathbb{E}[X_1] - \mu_1\mathbb{E}[X_2] + \mu_1\mu_2 = 0.
\end{aligned}
$$

Varying the mean of MVN inputs will result in shifted fANOVA components. Varying the variance of intput variables will not change the fANOVA decomposition and is therefore not investigated further.

Standard MVN, linear function, interaction, dependent inputs

So Figure 3 and Figure 4 are not how we want the fANOVA components to look like under dependence - but how do we want them to look like? Since we changed nothing about the structure of the function, should they generate identical plots as the classical components??
But on the other hand it cannot be that they give the same plots... I mean we wrote down the definition of the generalized components earlier, of course it is not the same as for the classical components. So as a function they literally look different. We couldn't compute them exactly but we could at least write out the system of equations we have to solve/ the probem that needs to be solved.
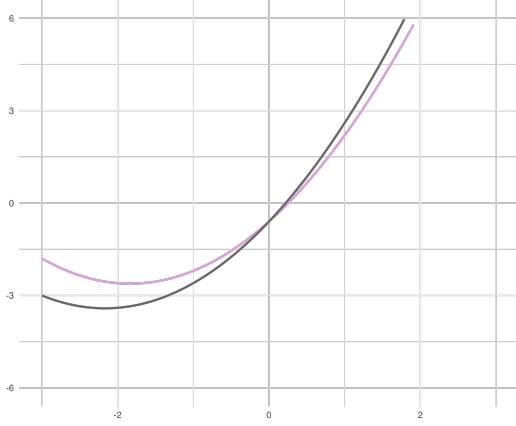
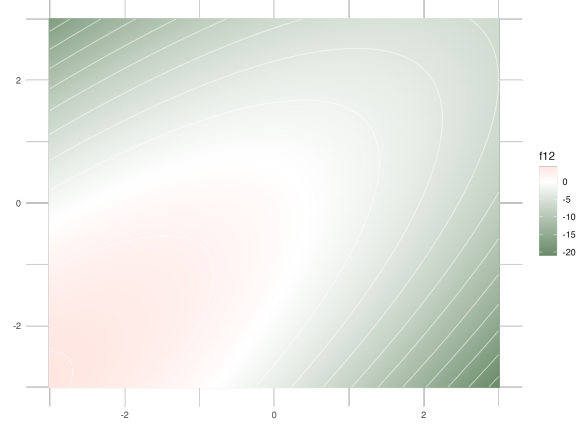Figure 3: Main terms as calculated via classical fANOVA for $g(x) = x_1 + 2x_2 + x_1 x_2$ with $\rho = 0.6$.

Figure 4: Contour plot of $g(x) = x_1 + 2x_2 + x_1 x_2$ with $\rho = 0.6$.

Uniform, quadratic, no interaction

Let us also try out another function $g(x_1, x_2) = a + x_1 + x_2^2$; to see how fANOVA deals with quadratic main effects. We will also change the distribution of the inputs. Lets us consider two independent random variables with uniform distribution over the interval $[-1, 1]$, meaning they are already centred. We calculate:

$$y_\emptyset = \mathbb{E}[g(X_1, X_2)] = a + \mathbb{E}[X_1] + \mathbb{E}[X_2^2] = a + 0 + \tfrac{1}{3} = a + \tfrac{1}{3}$$
$$y_1(x_1) = \mathbb{E}[g(x_1, X_2)] - y_\emptyset = a + x_1 + \tfrac{1}{3} - \left(a + \tfrac{1}{3}\right) = x_1$$
$$y_2(x_2) = \mathbb{E}[g(X_1, x_2)] - y_\emptyset = a + 0 + x_2^2 - \left(a + \tfrac{1}{3}\right) = x_2^2 - \tfrac{1}{3}$$
$$y_{1,2}(x_1, x_2) = g(x_1, x_2) - y_\emptyset - y_1(x_1) - y_2(x_2) = a + x_1 + x_2^2 - \left(a + \tfrac{1}{3} + x_1 + x_2^2 - \tfrac{1}{3}\right) = 0$$

The log fallacy Look again in Lengerich et al. (2020). fANOVA revealing that an interaction might not be an interaction after all, inside the log term is a product which suggests an interaction but it can be rewritten as a sum so maybe fANOVA will yield zero for the interaction?

Non-additive functions These functions demonstrate that fANOVA does not assume additive structure and can identify purely interactive effects. If variables only appear in interaction terms, their main effects vanish:

$$f_2(x) = x_1 \cdot x_2 \cdot x_3 \cdot x_4$$

Identifying interaction clusters This example shows that all the cross-terms vanish. So fANOVA correctly identifies that we have interaction pairs and no term is appearing multiple times in two or more interaction groups.

$$f(x) = (x_1 + x_2)^2 + x_3 x_4 + \sin(x_5 x_6)$$

$$y_\emptyset = \mathbb{E}[f] = \mathbb{E}[(X_1 + X_2)^2] + \mathbb{E}[X_3 X_4] + \mathbb{E}[\sin(X_5 X_6)] = 2 + 0 + 0 = 2$$

$$y_1(x_1) = \mathbb{E}_{X_{2,3,4,5,6}}[f(x)] - y_\emptyset = x_1^2 + 1 - 2 = x_1^2 - 1$$
$$y_2(x_2) = x_2^2 - 1$$
$$y_3(x_3) = \mathbb{E}[x_3 X_4] - y_\emptyset = 0$$
$$y_{12}(x_1, x_2) = \mathbb{E}[f(x)] - y_\emptyset - y_1(x_1) - y_2(x_2) = (x_1 + x_2)^2 - 2 - (x_1^2 - 1) - (x_2^2 - 1) = 2x_1 x_2$$
$$y_{13}(x_1, x_3) = \mathbb{E}_{X_{2,4,5,6}}[f(x)] - y_\emptyset - y_1(x_1) - y_3(x_3) = x_1^2 + 1 - 2 - (x_1^2 - 1) - 0 = 0$$
$$y_5(x_5) = \mathbb{E}_{X_6}[\sin(x_5 X_6)] - y_\emptyset = 0$$
$$y_6(x_6) = 0$$
$$y_{56}(x_5, x_6) = \sin(x_5 x_6)$$

Additive structure and necessity of interaction term $f(x_1, x_2, x_3) = x_1 + x_2 + x_3 + \epsilon x_1 x_2 x_3$ is mostly additive and has a minor interaction effect (depending on the value of ϵ), one can for example decide whether to leave in the interaction effect in a surrogate model or if it is negligible. We assume zero centred independent inputs, i.e. $X_i \sim \mathcal{N}(0, 1)$ for $i = 1, 2, 3$.

$$y_\emptyset = \mathbb{E}[f(X)] = 0$$
$$y_1(x_1) = \mathbb{E}_{X_2, X_3}[f(x_1, X_2, X_3)] - y_\emptyset = x_1$$
$$y_2(x_2) = x_2$$
$$y_3(x_3) = x_3$$
$$y_{12}(x_1, x_2) = y_{13}(x_1, x_3) = y_{23}(x_2, x_3) = 0$$
$$y_{123}(x_1, x_2, x_3) = f(x_1, x_2, x_3) - y_1(x_1) - y_2(x_2) - y_3(x_3) = \epsilon x_1 x_2 x_3$$
$$\mathrm{Var}[f] = \mathrm{Var}[x_1] + \mathrm{Var}[x_2] + \mathrm{Var}[x_3] + \mathrm{Var}[\epsilon x_1 x_2 x_3] = 3 + \epsilon^2$$
$$\mathrm{Var}[y_1] = \mathrm{Var}[y_2] = \mathrm{Var}[y_3] = 1$$
$$\mathrm{Var}[y_{123}] = \epsilon^2$$

8 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.

non-centred inputs: their effect terms are shifted, so that fANOVA components are centred around zero; other than that, shifting does not change nature/ form of the effects Generalization to dependent inputs: we were able to write out the problem but not find the solution; solve simultanesouly; analytically possible? Or only numerically? Outdated or unfinished packages, might hide in other packages, but difficult to identify. mlr3 function but then has to be part of the entire pipeline, need to train a model, and therefore cannot ensure that independence assumption holds.

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.

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

What is the final conclusion of my work? What do I even want to say with my thesis? fANOVA powerful theory, sound mathematical foundation, but without standardized software implementation application to IML difficult. But fANOVA is "applied" by researchers, how? For theoretical results?

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

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:

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

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

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

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

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).

Efron, B. and Stein, C. (1981). The Jackknife Estimate of Variance, **The Annals of Statistic**(Vol. 9, No. 3): pp. 586–596.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine., *The Annals of Statistics* **29**(5): 1189–1232. Publisher: Institute of Mathematical Statistics.
URL: *https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full*

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*

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*

Huang, J. Z. (1998a). Functional ANOVA Models for Generalized Regression, *Journal of Multivariate Analysis* **67**(1): 49–71.
URL: *https://ideas.repec.org/a/eee/jmvana/v67y1998i1p49-71.html*

Huang, J. Z. (1998b). Projection estimation in multiple regression with application to functional ANOVA models, *The Annals of Statistics* **26**(1).
URL: *https://projecteuclid.org/journals/annals-of-statistics/volume-26/issue-1/Projection-estimation-in-multiple-regression-with-application-to-functional-ANOVA/10.1214/aos/1030563984.full*

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*

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*

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*

Owen, A. (2003). The dimension distribution and quadrature test functions, *Statistica Sinica* **13**: 1–17.

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*

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*

Vaart, A. W. v. d. (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