Bachelor's Thesis

fANOVA for Interpretable Machine Learning

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, Month Dayth, 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

Questions

Clarifying the connection between fANOVA, expected value, and projection & bringing together different definitions

When I start from Muehlenstaedt et al. (2012) and basically go the way: *fANOVA term* $\rightarrow$ *expected value* $\rightarrow$ *projection*, I arrive at the formulation where we first compute the projection and afterwards subtract lower order terms. For example following the definition by Muehlenstaedt et al. (2012) and using the parallel between (conditional) expected value and the projection (Van Ravenzwaaij et al., 2018) we can write for example for $u = \{1, 2\}$:

$$\begin{aligned} y_{12} &:= \mathbb{E}[y(\boldsymbol{X}) \mid X_1 = x_1, X_2 = x_2] - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2) \\ &= \arg \min_{g_{\{1,2\}} \in \mathcal{G}_{\{1,2\}}} \mathbb{E}\left[(y(\boldsymbol{X}) - g_{\{1,2\}}(X_1, X_2))^2\right] - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2) \\ &= (\Pi_{\mathcal{G}_{\{1,2\}}} y(\boldsymbol{X})) - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2) \end{aligned}$$

But when I get it correctly Hooker (2007) writes his generalized fANOVA components as the projection of the differences. Also, he goes the other way around, starting from the projections (and we could restate this as the conditional expected value).
Because Hooker defines the fANOVA terms as:

$$\{f_u(x_u) \,|\, u \subseteq d\} = \arg \min_{\{g_u \in L^2(\mathbb{R}^u)\}_{u \subseteq d}} \int \left(y(\boldsymbol{x}) - \sum_{u \subseteq d} g_u(x_u) \right)^2 w(\boldsymbol{x}) \, d\boldsymbol{x} \tag{1}$$

And I think we can rewrite this as the conditional expected value. For example for $u = \{1, 2\}$:

$$\begin{aligned} y_{12}(.;.) &:= \arg \min_{g_{\{1,2\}} \in L^2(\mathbb{R}^2)} \int \left(y(\boldsymbol{x}) - \sum_{\{1,2\}} g_{\{1,2\}}(x_1, x_2) \right)^2 w(\boldsymbol{x}) \, d\boldsymbol{x} \\ &= \arg \min_{g_{\{1,2\}} \in L^2(\mathbb{R}^2)} \int \left(y(\boldsymbol{x}) - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2) - g_{\{1,2\}}(x_1, x_2) \right)^2 w(\boldsymbol{x}) \, d\boldsymbol{x} \\ &= \arg \min_{g_{\{1,2\}} \in L^2(\mathbb{R}^2)} \mathbb{E}[(y(\boldsymbol{X}) - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2) - g_{\{1,2\}}(x_1, x_2))^2] \\ &= \Pi_{\mathcal{G}_{\{1,2\}}}(y - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2)) \\ &= \mathbb{E}[y(\boldsymbol{X}) - y_\emptyset - y_{\{1\}}(x_1) - y_{\{2\}}(x_2) | X_1 = x_1, X_2 = x_2] \end{aligned}$$

So Hooker (2007) defines the fANOVA terms via the projection *fANOVA term* $\rightarrow$ *projec-*

tion → expected value (not in Hooker). But he takes the projection of the differences.

- Could it happen that based on fANOVA decomposition we build a model which uses the interaction effect of i, j but not the main effects i and/or j?
- Can I really compute the generalized fANOVA terms a proposed by Hooker (2007) by hand? Molnar writes: "The estimation is done on a grid of points in the feature space and is stated as a minimization problem that can be solved using regression techniques. However, the components cannot be computed independently of each other, nor hierarchically, but a complex system of equations involving other components has to be solved. The computation is therefore quite complex and computationally intensive." *If he would write he generalized fANOVA as conditional expected values it is actually not that complicated and we simply would need to solve regression problems hierarchically.*
- I am still confused if setting the zero-mean constraint for $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0, 1]$ is essentially saying that we centre the distribution and now assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-1, 1]$. So can we, instead of explicitly stating the zero-mean constraint just assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-1, 1]$? And following the same principle we would shift other distributions by altering their parameters, not by explicitly stating the zero-mean constraint? Then for the standard normal distribution we wouldn't need to do anything, it is already centred around 0. For other distributions we would need to change, and for some it doesn't make inhaltlichen Sinn e.g. Poisson distribution? *No generally zero-mean constraint and distribution assumption of the variables has no connection. The zero-mean constraint is sth. we set for the fANOVA terms y_u while distribution assumption is about the input variables X_i. We don't set the zero-mean constraint for the variables"*
-
- fANOVA decomposition via the integral, how would the zero mean constraint look here? (see "General_fANOVA_handnotes")
- Can you reconstruct the function from only the fANOVA terms? I think it can be reconstructed only if variables are independent, have zero-mean, are orthogonal?
- Is it possible to perform fANOVA for non-square-integrable functions? *in general yes but the variance decomposition doesn't work then or might have problems.*
- fANOVA decomposition for discrete variables possible? Does it make sense even?

- Connection between the (conditional) expected value, (partial) integral, projections (section **??**)?
- In the hierarchical orthogonality condition (4.2) formulated in Hooker (2007) for the gerneralized fAVNOA framework, shouldn't we explicitly exclude the case that $v = u$, because then, we would require that the inner product of the fANOVA component is zero wouldn't we (section 5)?
- Why is it a problem, when explainability methods also place large emphasis on regions of low probability mass when dependencies between variables exist - because in the end explainability is about explaining the model, not the data generating process; and after all it is how the model works in these regions. [But as the Hooker example illustrates, how the model works and what it estimates in these regions is wrong and then it's better to not report any model behaviour or come closer to the DGP than to give wrong estimations?]
-
- Use of AI tools?
- Do we need to restrict ourselves to the unit hypercube? Or does fANOVA decomposition work in general, but maybe with some constraints? Originally it was constructed for models on the unit hypercube $[0, 1]$, but other papers also use models from R^d *Generally no restriction, so next step could be to generalize, to* $\mathbb{R}^n$*, other measures, dependent variables*
- Still unclear: Are the terms fully orthogonal or hierarchically? See subsection on Orthogonality of the fANOVA terms (especially the example) I think in the original fANOVA decomposition the terms are orthogonal but in the generalized fANOVA (Hooker, 2007) they are hierarchically orthogonal. *fully orthogonal when independence assumption, probably partially when no independence*
- $x_1, \ldots, x_k$ are simply the standardized features, right? *Yes*
- My current understanding: we need independence of $x_1, \ldots x_k$ so that fANOVA decomposition is unique (and orthogonality holds). We need zero-mean constraint for the orthogonality of the components. We need orthogonality for the variance decomposition.*zero-mean* $\rightarrow$ *orthogonality* $\rightarrow$ *uniqueness; Lemma 1 im Hooker 2007 ist verallgemeinerg ds zero-mean constraint*

- Next step might be to investigate the (mathematical) parallels of fANOVA decomposition and other IML methods (PDP, ALE, SHAP), e.g. there is definitely a strong relationship between Partial dependence (PD) and fANOVA terms, and PD is itself again related to other IML methods; Also look how are other IML models studied and study fANOVA in a similar way (e.g. other IML methods are defined, checked for certain properties, examined under different conditions (dependent features, independent features) etc.) (see dissertation by Christoph Molnar for this); Also I would be very interested in investigating the game theory paper further (Fumagalli et al., 2025) but still a bit unsure if it is too complex.

- Why does a fANOVA decomposition of a simple GAM not lead to the "true" coefficients? https://christophm.github.io/interpretable-ml-book/decomposition.html talks about this a bit in the subchapter "Statistical regression models" *It should actually lead to the GAM; at least under all the constraint like zero-mean constraint and orthogonality*

-

- In Hooker (2004) they work with $F(x)$ and $f(x)$, but in Sobol (2001) they only work with $f(x)$. I think this is only notation? *Only notation.*

- Does orthogonality in fANOVA context mean that all terms are orthogonal to each other? Or that a term is orthogonal to all lower-order terms ("Hierarchical orthogonality ")? *The terms are hierarchically orthogonal, so each term is orthogonal to all lower-order terms, but not to the same-order terms! So f_1 is not necessarily orthogonal to f_2 but it is orthogonal to f_{12}, f_0.*

- Do the projections here serve as approximations? (linalg skript 2024 5.7.4 Projektionen als beste Annäherung) *Yes, they can be interpreted as sort of approximation.*

- Which sub-space are we exactly projecting onto? Are the projections orthogonal by construction (orthogonal projections) or only when the zero-mean constraint is set? *The subspace we project onto depends on the component. For f_0 we project onto the subspace of constant functions, for f_1 we project onto the subspace of all functions that involve x_1 and have an expected value of 0 (zero-mean constraint to ensure orthogonality). It depends on the formulation of the fANOVA decomposition if you need to explicitly set the zero-mean constraint for orthogonality or if it is met by construction.*

- How "far"should I go back, formally introduce L^2 space, etc. or assume that the reader is familiar with it? *Yes, space, the inner product on this space should be formally introduced.*

2 Background Knowledge

2.1 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 $f_{\boldsymbol{X}} : \mathbb{R}^N \to \mathbb{R}_0^+$.
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:

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

The norm is then defined as:

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

Which resource should I cite for these "general" definitions? e.g. https://apachepersonal.miun.se/ andrli/Bok.pdf?

2.2 Conditional expectation

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 $X_1 = x_1$:

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

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 $\mathcal{L}^2$:

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

I think this is closely related to Hilbert projection theorem?
Definition of $\mathcal{L}^2$ space and projection modified from https://tnagler.github.io/mathstat-lmu-2024.pdf.

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 $X_i \perp X_j$. Lastly, for any real vector $\boldsymbol{a} \in \mathbb{R}^d$, the linear combination $\boldsymbol{a}^\top \boldsymbol{X}$ is normally distributed:

$$\boldsymbol{a}^\top \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}).$$

Find official resource for these properties; review how this can be shown.

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 famous paper 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. 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". Sobol indices are now commonly used in sensitivity analysis. Efron and Stein (1981) use the idea of the decomposition to proof their famous lemma on jackknife variances. 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 example citations needed.

3.2 Modern Work on fANOVA

The fANOVA decomposition has a long history with roots in mathematical statistics and non-parametric estimation theory. 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. 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.
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). 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).
Liu and Owen (2006) use of fANOVA and sensitivity analysis for functions arising in computational finance. 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).

fANOVA and U-statistics, fANOVA and sensitivity analysis, fANOVA and GAMs (with interactions)

4 Classical fANOVA

4.1 Formal Introduction to fANOVA

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 $y(\boldsymbol{X})$ be 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:*

$$y(\boldsymbol{X}) = \sum_{u \subseteq \{1,\ldots,N\}} y_u(\boldsymbol{X}_u), \tag{2}$$

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.
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:*

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

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:

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

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 3, Equation 4)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:

$$\int y_u(\boldsymbol{x}_u) f_{X_i}(x_i)\, d\nu(x_i) = 0, \quad \text{for } i \in u \neq \emptyset. \tag{5}$$

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

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

One can follow the same reasoning for the orthogonality condition:

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

Construction of 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:

$$y_\emptyset = \int y(\boldsymbol{x}) \prod_{i=1}^{N} f_{X_i}(x_i)\, d\nu(x_i) = \mathbb{E}[y(\boldsymbol{X})]. \tag{6}$$

For all other effects $u \neq \emptyset$ we can write:

$$y_u(\boldsymbol{X}_u) = \int y(\boldsymbol{X}_u, \boldsymbol{x}_u) \prod_{i=1, i \notin u}^{N} f_{X_i}(x_i)\, d\nu(x_i) - \sum_{v \subsetneq u} y_v(\boldsymbol{X}_v), \tag{7}$$

The basis 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.

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 $y_{i_1,\ldots,y_n} \in \mathcal{L}^2$ 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:

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

The variance of the fANOVA components is then defined as

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

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

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

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

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

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:

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

The main effect $y_i(x_i)$ is the projection of y onto the subspace of all functions that only depend on x_i and have an expected value of zero while accounting for the lower-order effects. The subspace we project onto is $\mathcal{G}_i = \{g(x) = g_i(x_i); \int g(x) d\nu(x_i) = 0\}$. 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):

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

The two-way interaction effect $y_{ij}(x_i, x_j)$ is the projection of y onto the subspace of all functions that depend on x_i and x_j and have an expected value of zero in each of it's single components, i.e. $\mathcal{G}_{i,j} = \{g(x) = g_{ij}(x_i, x_j); \int g(x) d\nu(x_i) = 0 \wedge \int g(x) d\nu(x_j) = 0\}$. Again, we account for lower-order effects by subtracting the constant term and all main effects:

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

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

$\{g(\boldsymbol{x}) = g_u(\boldsymbol{x}_u); \int g(\boldsymbol{x}) d\nu(\boldsymbol{x}_u) = 0\}$:

$$
\begin{aligned}
(\Pi_{\mathcal{G}_u} y) - \sum_{v \subsetneq u} y_v(x) &= \arg\min_{g_u \in \mathcal{G}_u} \|y - g_u\|^2 - y_0 \\
&= \arg\min_{g_u \in \mathcal{G}_u} \mathbb{E}[(y(\boldsymbol{X}) - g(X_u))^2] \\
&= \mathbb{E}[y(\boldsymbol{X}) | X_u = x_u] - \sum_{v \subsetneq u} y_v(x) = y_u(\boldsymbol{x}_u),
\end{aligned}
$$

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.

Notes & Questions

Situation: $y(\boldsymbol{X}) \in \Omega, \mathcal{G} \subseteq \Omega, g(\boldsymbol{X}) \in \mathcal{G}$.
Vaart (1998) tells us that the expected value is equivalent to the projection Muehlenstaedt et al. (2012) tells us that the fANOVA terms are equivalent to the conditional expected value.

5 Generalized fANOVA

5.1 Motivating Example

Some good example that illustrates how dependent input variables can lead to misleading results and break the fANOVA.

5.2 Formal Introduction to Generalized fANOVA

In practice assuming independent input variables is often not realistic. When this assumption is violated, it can lead to misleading results (Hooker, 2007). The need for a generalization to dependent variables is evident. We base this chapter mainly on the generalization of Rahman (2014), while there exists other work from Hooker (2007) or Chastaing et al. (2012).
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 $u \subseteq d$.
The definition of the fANOVA decomposition, and the two main properties - zero mean and orthogonality - can be stated in the same way as for the classical case, with s slight change for the orthogonality. We build now on the weak annihilating conditions instead of the strong ones, which results in hierarchical orthogonality instead of full orthogonality.

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

$$y(\boldsymbol{X}) = \sum_{u \subseteq \{1,\dots,N\}} y_{u,G}(\boldsymbol{X}_u) \tag{8}$$

The subscript G indicates that we are working with the generalized fANOVA components. The rest is equivalent to the classical fANOVA. The difference comes from how the components look, which we will see in the following.

Proposition 5.1. ***Weak annihilating conditions.*** *The weak annihilating conditions are the foundation for the generalized fANOVA decomposition. They are exactly formulated as the strong annihilating conditions, with the exception that the integral uses the joint-probability density function of the variables of interest, instead of the individual marginal probability density functions. This makes sense, since for the general case we cannot ensure to recover the marginal densities from the joint density function. Is this really the reason? Or is the reason: When there are dependencies between variables then*

the individual pdfs would not assign the "correct weight" as they ignore the dependence between features in u. The weak annihilating conditions require that:

$$\int_{\mathbb{R}} y_{u,G}(\boldsymbol{x}_u) f_{\boldsymbol{X}_u}(\boldsymbol{x}_u) d\nu(x_i) = 0 \quad \textit{fori} \in u \neq \emptyset \tag{9}$$

Proposition 5.2.

Proposition 5.3.

5.3 Generalization by Hooker

Again multiple work on generalization of fANOVA, e.g. Hooker (2007), Rahman (2014), Chastaing et al. (2012). We want to let go of two key assumptions of the classical fANOVA decomposition (as introduced by Sobol (1993)): We widen the input domain to the multidimensional reel number line, i.e. we now work in the measure space $(X, \mathcal{F}, \nu) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), dw(x))$. This goes hand in hand with dropping the assumption about the uniform distribution of the X_i. Further, we investigate what happens when the variables are no longer independent of each other.
The inner product on $\mathcal{L}^2(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), dw(x))$ is now defined more generally as the integral of a weighted product:

$$\langle f, g \rangle = \int f(x) g(x)\, d\nu(x) \quad \forall f, g \in \mathcal{L}^2 \quad \text{with} \quad \nu(dx) = w(x)dx$$

The norm is given by

$$\|f\|_w = \sqrt{\langle f, f \rangle_w} = \sqrt{\int f^2(x)\, w(x) dx} \quad \forall f \in \mathcal{L}^2$$

The general definition of the function $f(x)$ as a weighted sum stays the same (see Equation 2). What changes is the definition of the fANOVA components. The components are simultaneously defined as:

$$\{f_u(x_u) \,|\, u \subseteq d\} = \arg \min_{\{g_u \in L^2(\mathbb{R}^u)\}_{u \subseteq d}} \int \left(\sum_{u \subseteq d} g_u(x_u) - f(x) \right)^2 w(x)\, dx \tag{10}$$

There is a key difference to the classical definition: All the components are defined simultaneously via the orthogonal projections of the original function $f(x)$. This means the components f_u are a set of functions that jointly minimize the weighted squared difference to the original function $f(x)$ and fulfil the generalized zero-mean constraint and hierarchi-

cal orthogonality (both defined in the following). A natural choice for the weights $w(x)$ is the probability distribution of the x_i (Hooker, 2007).
We require the fANOVA terms to be centred around the grand mean, in the same way as we did for the classical approach. Hooker (2007) formulates this in a generalized zero-mean condition for dependent variables:

$$\forall u \subseteq d,\ \forall i \in u: \quad \int f_u(x_u)\, w(x)\, dx_i\, dx_{-u} = 0 \tag{11}$$

Orthogonality of the fANOVA terms plays and important role. It ensures that they represent isolated effects which makes the interpretation of fANOVA so useful in practice. In contrast to the classical fANOVA, we set a hierarchical orthogonality constraint (instead of a general orthogonality constraint):

$$\forall v \subseteq u,\ \forall g: \quad \int f_u(x_u)\ g_v(x_v) w(x)\ dx = 0 \tag{12}$$

I am always puzzled by this definition because v could theoretically be equal to u which would require the function to be orthogonal to itself. But wanting this for all functions g somehow changes something, but I am not super clear why. Would it be correct to write:

$$\forall v \subset u: \quad \int f_u(x_u)\ g_v(x_v) w(x)\ dx = 0 \tag{13}$$

Category	**Classical**	**Generalized**
Measure space	$([0,1]^n, \mathcal{B}([0,1]^n))$	$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$
Measure	$\mathbb{P} : \mathcal{B}([0,1]^n) \to [0,1]$	$\mu : \mathcal{B}(\mathbb{R}^n) \to [0,\infty)$, where $\mu(A) = \int_A w(x)\,dx$, $w(x) = \frac{d\mu}{d\lambda}$
Distribution assumption	$\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} \mathcal{U}([0,1])$	$\mathbf{X} = (X_1, \ldots, X_n) \sim$ any distribution
Random Variable	$\mathbf{X} : \Omega \to [0,1]^n$, $\mu := \mathbf{X}_{\#}\mathbb{P}$	$\mathbf{X}_* : \Omega \to \mathbb{R}^n$, $w(x)\,dx = \mathbf{X}_{\#}\mathbb{P}$
Inner product	$\langle f, g\rangle = \int f(x)g(x)\,dx$	$\langle f, g\rangle_w = \int f(x)g(x)\,w(x)\,dx$
Norm	$\|f\| = \left(\int f(x)^2\,dx\right)^{1/2} = \sqrt{\mathbb{E}[f(\mathbf{X})^2]}$	$\|f\|_w = \left(\int f(x)^2 w(x)\,dx\right)^{1/2} = \sqrt{\mathbb{E}[f(\mathbf{X})^2]}$
fANOVA components	$f_u(x) = \int_{x_{-u}} \left(F(x) - \sum_{v\subset u} f_v(x)\right) dx_{-u}$	$\{f_u(x_u)\}_{u\subset d} = \arg\min_{\{g_u \in L^2(\mathbb{R}^u)\}} \int \left(\sum_{u\subset d} g_u(x_u) - F(x)\right)^2 w(x)\,dx$
Zero-mean constraint	$\int f_u(x_u)\,dx_u = 0$ for $u \neq \emptyset$	$\forall u \subset d, \forall i \in u : \int f_u(x_u)\,w(x)\,dx_i\,dx_{-u} = 0$
Orthogonality	$\int f_u(x_u) f_v(x_v)\,dx = 0$ for $u \neq v$	$\forall v \subset u,\ \forall g_v : \int f_u(x_u) g_v(x_v) w(x)\,dx = 0$

Table 1: Comparison of classical and generalized functional ANOVA (fANOVA) decompositions.

6 Estimation of fANOVA

In this chapter we will illustrate two approaches to estimate the fANOVA components on a conceptual level. The first approach can be found in Hooker (2007) and is essentially a linear least squares problem. The second approach uses Fourier polynomial expansion and is proposed by Rahman (2014).

7 Examples

7.1 Multivariate Normal Inputs

Consider the function $g = a + X_1 + 2X_2 + X_1X_2$. We assume that $\boldsymbol{X} = (X_1, X_2)^T$ follows a MVN distribution:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

where we assume that $\sigma_1 = \sigma_2 = 1$, while we will vary the values for ρ_{12} to distinguish between varying degrees of dependence between the two variables. From the properties

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

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

We also know that the conditional distributions are:

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}(\rho_{12} x_2, 1 - \rho_{12}^2), \quad X_2 \mid X_1 = x_1 \sim \mathcal{N}(\rho_{12} x_1, 1 - \rho_{12}^2)$$

Case 1: Independent Inputs

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 instead of the integral for notational simplicity:

$$f_0 = \mathbb{E}[g_1(X_1, X_2)] = \mathbb{E}[a + X_1 + 2X_2 + X_1 X_2] = \mathbb{E}[a] + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1 X_2]$$

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:

$$f_0 = \mathbb{E}[a] + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1]\mathbb{E}[X_2] = a$$

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

$$\begin{aligned}
f_1(x_1) &= \mathbb{E}_{X_2}[g_1(x_1, X_2)] - f_0 \\
&= \mathbb{E}_{X_2}[a + x_1 + 2X_2 + x_1 X_2] - a \\
&= x_1 + 2\mathbb{E}[X_2] + x_1 \mathbb{E}[X_2] = x_1 \\
f_2(x_2) &= \mathbb{E}_{X_1}[g_1(X_1, x_2)] - f_0 \\
&= \mathbb{E}_{X_1}[a + X_1 + 2x_2 + X_1 x_2] - a \\
&= \mathbb{E}_{X_1}[X_1] + 2x_2 + x_2 \mathbb{E}_{X_1}[X_1] = 2x_2 \\
f_{12}(x_1, x_2) &= \mathbb{E}[g_1(x_1, x_2)] - f_0 - f_1(x_1) - f_2(x_2) \\
&= a + x_1 + 2x_2 + x_1 x_2 - a - x_1 - 2x_2 = x_1 x_2
\end{aligned}$$

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. In case of independence the generalized fANOVA decomposition is of course equal to the classical one. We show this simple example nevertheless to illustrate at which step

which assumption is used. This motivates the need for a generalization of the fANOVA decomposition to dependent variables.

Case 2: Dependent Inputs (weak)

Now assume $\rho_{12} = 0.2$, $\sigma_1 = \sigma_2 = 1$. The constant term f_0 is given by:

$$\begin{aligned} f_0 &= \mathbb{E}[g(X_1, X_2)] = a + \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_1 X_2] \\ &= a + \mathbb{E}[X_1 X_2] = a + (\mathrm{Cov}(X_1, X_2) + \mathbb{E}[X_1]\mathbb{E}[X_2]) \\ &= a + \rho_{12} \end{aligned}$$

The main effects can be computed as:

$$\begin{aligned} f_1(x_1) &= \mathbb{E}[g(X_1, X_2)|X_1 = x_1] - f_0 \\ &= \mathbb{E}[a + x_1 + 2X_2 + x_1 X_2 | X_1 = x1] - (a + \rho_{12}) \\ &= a + x_1 + 2\mathbb{E}[X_2|X_1 = x_1] + x_1\mathbb{E}[X_2|X_1 = x_1] - a - \rho_{12} \\ &= x_1 - \rho_{12} + \rho_{12}(2 + x_1) \\ f_2(x_2) &= \mathbb{E}[g(X_1, X_2) \mid X_2 = x_2] - f_0 \\ &= \mathbb{E}[a + X_1 + 2x_2 + X_1 x_2 \mid X_2 = x_2] - (a + \rho_{12}) \\ &= a + 2x_2 + x_2\mathbb{E}[X_1 \mid X_2 = x_2] - a - \rho_{12} \\ &= 2x_2 + x_2^2\rho_{12} + \rho_{12} \end{aligned}$$

Finally, the interaction effect $f_{12}(x_1, x_2)$ is given by:

$$\begin{aligned} f_{12}(x_1, x_2) &= g(x_1, x_2) - f_0 - f_1(x_1) - f_2(x_2) \\ &= a + x_1 + 2x_2 + x_1 x_2 - (a + \rho_{12}) - (x_1 - \rho_{12} + 2\rho_{12}x_1 + \rho_{12}x_1^2) - (2x_2 + x_2^2\rho_{12} + \rho_{12}) \\ &= x_1 x_2 - 2\rho_{12}x_1 - \rho_{12}x_1^2 - \rho_{12}x_2^2 - \rho_{12} \end{aligned}$$

Case 3: Dependent Inputs (strong)

8 Software Implementation

8.1 Software implementations

- Suitable but currently problems installing locally: fanova
- Context of kriging models; create own graphs (not super informative): fanovaGraph

- mlr3 function
- tntorch
- shapley values implementation python

9 Conclusion

10 Mathematical Statements

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

A Appendix

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*

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*

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.

Van Ravenzwaaij, D., Cassey, P. and Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling, *Psychonomic Bulletin & Review* **25**(1): 143–154.
URL: *http://link.springer.com/10.3758/s13423-016-1015-8*

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