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

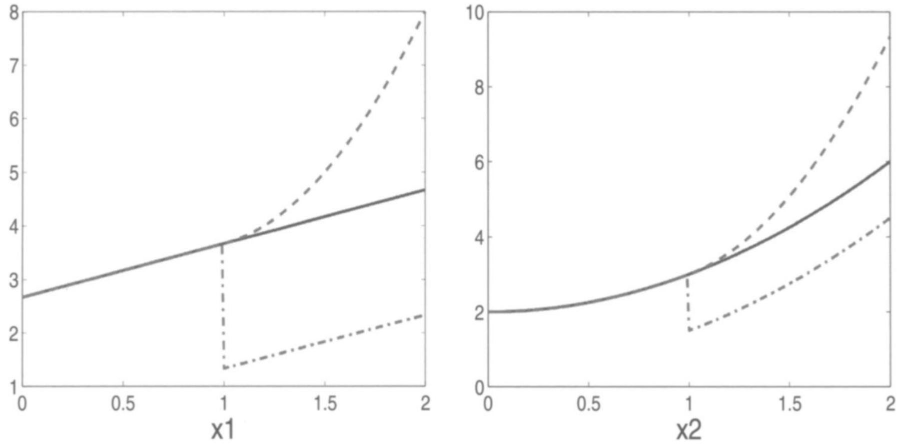

Figure 2. A comparison of Functional ANOVA effects between $F(x_1, x_2)$ (solid) and a learned approximation, $\hat{F}(x_1, x_2)$ (dashed). These two functions are indistinguishable on the data in Figure 1. The left-hand plot provides the effect for x_1, the right for x_2. Dotted lines provide the conditional expectations $E(\hat{F}(x_1, x_2)|x_1)$ and $E(\hat{F}(x_1, x_2)|x_2)$ with respect to the measure in Figure 1.

Questions

- How is the generalized fANOVA really computed? Could you compute it by hand?
- Effect of non-linear main effects in classical fANOVA (see "General_fANOVA_handnote")
- 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? I think 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 2)?
- 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 4)?

- I am a bit confused by Figure 2 in Hooker (2007) (see section 1), especially by the dotted line for conditional expectation. What should it tell us? I think what the dashed line (learned approximation) shows is that the model estimates a non-linear effect for x1, even though the true effect is linear. The reasons are the problematic data points in the top right region.
- 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 General Definitions

2.1 $\mathcal{L}_2$ space

Let $(X, \mathcal{F}, \nu)$ be a measure space, where X is a sample space, $\mathcal{F}$ is a $\sigma-$algebra for X and ν is a general measure. Then the vector space of all square-integrable functions is given by

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

$\mathcal{L}^2$ 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?

Orthogonal projection

$\mathcal{G} \subset \mathcal{L}^2$ denotes 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.

3 Foundations

3.1 Early Work on fANOVA

Hoeffding decomposition 1948

- The idea of fANOVA decomposition dates back to Hoeffding (1948).
- Introduces Hoeffding decomposition (or U-statistics ANOVA decomposition).
- Math-workings: involves orthogonal sums, projection functions, orthogonal kernels, and subtracting lower-order contributions.
- Assupmtions: unclear about all but one assumptions is (mututal?) independence of input variables, which is unrealistic in practice (different generalizations to dependent variables follow, e.g. Il Idrissi et al. (2025))
- Relevance: shows that U-statistics or any symmetric function of the data can be broken down into simpler pieces (e.g., main effects, two-way interactions) without overlap.
- Pieces can be used to dissect/explain the variance.
- fANOVA performs a similar decomposition, not for U-statistics but for functions.

⇒fANOVA and U-statistics

Sobol Indices 1993, 2001

- In "Sensitivity Estimates for Nonlinear Mathematical Model" (1993), Sobol first introduces decomposition into summands of different dimensions of a (square) integrable function.
- Does not cite Hoeffding nor discuss U-statistics.
- "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimate" (2001) builds on his prior work (Sobol, 2001).
- Math-workings: similar to Hoeffding, involving orthogonal projections, sums, and independent terms.
- Sobol focuses on sensitivity analysis for deterministic models, while Hoeffding is concerned with estimates of probabilistic models.

I think in his 1993 paper Sobol mainly introduces fANOVA decomposition (definition, orthogonality, L1 integrability), already speaks of L2 integrability and variance decompsoition, which leads to Sobol indices, gives some analytical examples and MC algorithm for calculations. In the 2001 paper he focuses on illustrating three usecases of the sobol indices + the decomposition

- ranking of variables
- fixing unessential variables
- deleting high order members

For each of the three there are some mathematical statements, sometimes an algorithm or an example. ⇒ **fANOVA and sensitivity analysis**

Efron and Stein (1981)

- Use idea to proof a famous lemma on jackknife variances (Efron and Stein, 1981)

Stone 1994

- Stone (1994)
- Math-workings: sum of main terms, lower-order terms, etc., with an identifiability constraint (zero-sum constraint); follows the same principle as the decomposition frameworks by Hoeffding (1948) and Sobol (2001).
- All of them work independently, do not cite each other, and use the principle with different goals/build different tools on it.
- Stone's work is part of a broader body of fANOVA models.

⇒fANOVA and smooth regression models / GAMs

I think the main focus of this paper is to extend the theoretical framework of GAMs with interactions. So the baseline is logistic regression with smooth terms but only univariate components are considered. Now the paper goes deeper into the theory where multivariate terms are also considered. For this they refer to the "ANOVA decomposition"of a function. The focus of the paper is on how the smooth multivariate interaction terms can be estimated, what mathematical properties they have, etc.

3.2 Modern Work on fANOVA

- Rabitz and Aliş (1999) see ANOVA decomposition as a specific high dimensional model representation (HDMR); the goal is to decompose the model iteratively from main effects, to lower order interactions and so on, but to do this in an efficient way and seletect only interaction terms that are necessary (most often lower-order interactions are sufficient). → chemistry paper

- Work of Hooker (2007) can be seen as an attempt to generalize Hoeffding decomposition (or the Hoeffding principle) to dependent variables. According to Slides to talk on Shapley and Sobol indices

- At least in his talk which is based on the paper Il Idrissi et al. (2025) he puts his work in a broader context of modern attempts to generalize Hoeffding indices. So Il Idrissi et al. (2025) can be seen as one attempt to generalize Hoeffding decomposition to dependent variables.

3.3 Formal Introduction to fANOVA

fANOVA decomposition

This chapter is based on the formal introductions by Sobol (1993, 2001), Hooker (2004), Owen, Muehlenstaedt et al. (2012). Where suitable we show both formulations of the fANOVA, via the integral and via the expected value. Let $i_1, \ldots, i_s$ denote a set of indices. For now, we assume that $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$ and work in the measure space $(X, \mathcal{F}, \nu) = ([0,1]^n, \mathcal{B}([0,1]^n), \lambda_n)$. $\mathcal{B}([0,1]^n)$ is the Borel σ-algebra on the n-dimensional unit interval and λ_n is the n-dimensional Lebesgue measure. The general inner product and norm we defined earlier simplify under these assumptions.
The inner product under uniform distribution assumption:

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

The norm under uniform distribution assumption:

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int f^2(x)\,d(x)} \quad \forall f \in \mathcal{L}^2$$

Definition. Let $f(x)$ be a mathematical model with input X_i as described above. We can represent such a model f as a sum of specific basis functions

$$f(x) = f_0 + \sum_{s=1}^{n} \sum_{i_1<...<i_s}^{n} f_{i_1...i_s}(x_{i_1}, ..., x_{i_s}) \tag{1}$$

To ensure identifiability and interpretation, we set the zero-mean constraint. It requires 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:

$$\int f_{i_1...i_s}(x_{i_1}, ..., x_{i_s})\, d\nu(x_k) = 0 \quad \forall k = i_1, ..., i_s \tag{2}$$

In combination with the zero-mean constraint (Equation 2), Sobol (1993) calls Equation 1 initially the “Expansion into Summands of Different Dimensions”. In Sobol (2001) he renames the decomposition to the “ANOVA-representation”. Now, it is mostly referred to as the “functional ANOVA decomposition”(Hooker, 2004).

The individual terms that make up Equation 1 are defined in the following. To get the constant term, we take the integral of f w.r.t. all variables:

$$f_0(x) = \int f(x) d\nu(x) = \mathbb{E}[f(X)] \tag{3}$$

The constant term f_0 captures the overall mean of f and serves as a baseline. Since the the remaining effects are centred around zero, they quantify the deviation from the overall mean. Next, we take the integral of y w.r.t. all variables except for x_i. This represents f as the sum of the constant term and the isolated effect of one variable x_i (main effect of x_i). This partial integral is equivalent to the expected value conditioned on the variable of interest x_i.

$$f_0 + f_i(x_i) = \int f(x) \prod_{k \neq i} \nu(d_{x_k}) = \mathbb{E}[f(X)|X_i = x_i] \tag{4}$$

Following the same principle, we can take the integral of f w.r.t. all variables except for x_i and x_j. With this we capture everything up to the interaction effect of x_i and x_j. This is equivalent to the expected value conditioned on both variables x_i and x_j:

$$f_0 + f_i(x_i) + f_j(x_j) + f_{ij}(x_i, x_j) = \int f(x) \prod_{k \neq i,j} \nu(d_{x_k}) = \mathbb{E}[f(X)|X_i = x_i, X_j = x_j] \tag{5}$$

For a successive construction of the fANOVA decomposition, we can generally write:

$$\int f(x) \prod_{k \notin u} \nu(d_{x_k}) = \mathbb{E}[f(X)|X_u = x_u] \tag{6}$$

With these partial integrations (or conditional expected values) we build up the fANOVA decomposition in a cumulative way. To actually see the fANOVA terms defined in isolation, it is clearer to rearrange terms. When we rearrange Equation 4 we see that the main effect of x_i is calculated by taking the marginal effect while explicitly accounting for what was already explained by lower-order terms, in this case the intercept:

$$f_i(x_i) = \int f(x) \prod_{k \neq i} \nu(d_{x_k}) - f_0 \tag{7}$$

The two-way interactions can then be seen as the marginal effects of the involved variables, while accounting for all main effects and the constant term:

$$f_{ij}(x_i, x_j) = \int f(x) \prod_{k \neq i,j} \nu(d_{x_k}) - f_0 - f_i(x_i) - f_j(x_j) \tag{8}$$

Therefore, it is also common to formulate the fANOVA decomposition in the following way (Hooker, 2007, 2004):

$$f_u(x) = \int_{[0,1]^{d-|u|}} \left(f(x) - \sum_{v \subsetneq u} f_v(x) \right) d\nu(x_{-u}). \tag{9}$$

This means we subtract all lower-order terms from the original function f and then integrate over all the variables not in u to get the effect of x_u. Using the linearity of the integral, we can also first take the partial integral of the original function w.r.t. all variables not in u and then subtract all the lower-order terms, as we did above for the main effects and two-way interaction effects. So generally we write:

$$f_u(x) = \int_{[0,1]^{d-|u|}} f(x) \, d\nu(x_{-u}) - \sum_{v \subsetneq u} f_v(x). \tag{10}$$

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.

Example 1

Before moving to properties of the fANOVA decomposition, let us introduce a simple function g as running example. It contains a constant term a, isolated linear effects of two variables x_1 and x_2 and their interaction.

$$g_1(x_1, x_2) = a + x_1 + 2x_2 + x_1x_2 \quad \text{for} \quad a, x_1, x_2 \in \mathbb{R}$$

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:

$$f_0 = \mathbb{E}[g_1(x_1, x_2)] = \mathbb{E}[a + x_1 + 2x_2 + x_1x_2] = \mathbb{E}[a] + \mathbb{E}[x_1] + 2\mathbb{E}[x_2] + \mathbb{E}[x_1x_2]$$

Making use of the independence assumption of x_1 and x_2, the last term can be written as the product of the expected values. Additionally, given the zero-mean 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:

$$\begin{aligned}
f_1(x_1) &= \mathbb{E}_{X_2}[g_1(x_1, X_2)] - f_0 \\
&= \mathbb{E}_{X_2}[a + x_1 + 2x_2 + x_1x_2] - a \\
&= x_1 + 2\mathbb{E}[x_2] + x_1\mathbb{E}[x_2] = x_1 \\
f_2(x_2) &= \mathbb{E}_{X_1}[g_1(X_1, x_2)] - f_0 \\
&= \mathbb{E}_{X_1}[a + x_1 + 2x_2 + x_1x_2] - a \\
&= \mathbb{E}_{X_1}[x_1] + 2x_2 + x_2\mathbb{E}_{X_1}[x_1] = 2x_2 \\
f_{12}(x_1, x_2) &= \mathbb{E}[g_1(x_1, x_2)] - f_0 - f_1(x_1) - f_2(x_2) \\
&= a + x_1 + 2x_2 + x_1x_2 - a - x_1 - 2x_2 = x_1x_2
\end{aligned}$$

It comes as no surprise that in this simple case the fANOVA decomposition does not provide any additional insights. This is because the model consists of only linear terms, constant terms, and an interaction. We show this simple example nevertheless to illustrate at which step we use which assumption. Understanding this will be relevant for the generalization of the method to dependent inputs later on. Also, it is interesting to compare this example with only linear effects (and an interaction) to the following, which will include a non-linear effect.

Example 2

We now look at the function $g_2 = a + x_1 + x_2^2$ which includes a quadratic effects. The constant fANOVA term is given by:

$$f_0 = \mathbb{E}[g_2(x_1, x_2)] = \mathbb{E}[a + x_1 + x_2^2] = a + \mathbb{E}[X_1] + \mathbb{E}[X_2^2] = a + \frac{1}{12}$$

This works because we are still in the setting, in which we assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$, in combination with the zero-mean constraint this allows us to state:

$$\mathbb{E}[X_2^2] = \mathbb{V}[X_2] = \frac{1}{12}(1-0)^2 = \frac{1}{12}$$

Next, we write the main effects:

$$f_1(x_1) = \mathbb{E}_{X_2}[g_2(x_1, X_2)] - f_0 = \mathbb{E}_{X_2}[a + x_1 + X_2^2] - f_0$$
$$= a + x_1 + \mathbb{E}[X_2^2] - f_0 = a + x_1 + \frac{1}{12} - \left(a + \frac{1}{12}\right) = x_1$$
$$f_2(x_2) = \mathbb{E}_{X_1}[g_2(X_1, x_2)] - f_0 = \mathbb{E}_{X_1}[a + X_1 + x_2^2] - f_0$$
$$= a + \mathbb{E}[X_1] + x_2^2 - f_0 = a + x_2^2 - \left(a + \frac{1}{12}\right) = x_2^2 - \frac{1}{12}$$

Finally, we compute the interaction effect:

$$f_{12}(x_1, x_2) = \mathbb{E}[g_2(x_1, x_2)] - f_0 - f_1(x_1) - f_2(x_2)$$
$$= a + x_1 + x_2^2 - \left(a + \frac{1}{12}\right) - x_1 - \left(x_2^2 - \frac{1}{12}\right) = 0$$

This example reveals something interesting about quadratic effects. It shows that they contain a constant effects, which is attributed to the constant terms f_0 and subtracted from the main effects. The quadratic effects does not influence the effects of the other terms, as the main effects of x_1 behaves as one would expect for a linear term; same goes for the non-existent interaction effect. Under the uniform distribution assumption, we can continue the calculations for higher powers of X_2 and still get a nice representation of the components, because following Cavalieri's quadrature formula we can find the moment of a uniformly distributed random variable X on the unit interval raised to the power of k as follows:

$$\mathbb{E}[X^k] = \int_0^1 x^k \, dx = \frac{1}{k+1} \quad \text{for } k \in \mathbb{N}_0 \tag{11}$$

Therefore, if we deal with a cubic term $g_3 = a + x_1 + x_2^2 + x_2^3$, we can compute the fANOVA decomposition as follows:

$$
\begin{aligned}
&f_0 = a + \mathbb{E}[X_1] + \mathbb{E}[X_2^2] + \mathbb{E}[X_2^3] = a + \frac{1}{12} + \frac{1}{4} = a + \frac{1}{3} \\
&f_1(x_1) = x_1 \\
&f_2(x_2) = x_2^2 - \frac{1}{12} + x_2^3 - \frac{1}{4} = x_2^2 + x_2^3 - \frac{7}{12} \\
&f_{12}(x_1, x_2) = 0
\end{aligned}
$$

Orthogonality of the fANOVA terms

Orthogonality of the fANOVA terms follows using the zero-mean constraint (Equation 2). If two sets of indices are not completely equivalent $(i_1, \ldots, i_s) \neq (j_1, \ldots, j_l)$ then

$$\int f_{i_1,\ldots,i_s} f_{j_1,\ldots,j_l} d(x) = 0 \tag{12}$$

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. In our example from before we can test this for $i = 1$ and $j = 2$:

$$\int f_1(x_1) f_2(x_2) d(x) = \int x_1 \cdot 2x_2 \, dx_1 \, dx_2 = \mathbb{E}[x_1 2x_2] = \mathbb{E}[x_1] \cdot 2\mathbb{E}[x_2] = 0$$

To write the expected value of a product as the product of the expected values we needed the independence assumption. To state that the product of the expected values is equal to zero, we used the zero-mean constraint. This shows that the independence assumption and zero-mean constraint are critical to ensure orthogonality in this traditional formulation of the fANOVA decomposition. This is of course also true for terms of different order, e.g. $f_{1,2}(x_1, x_2)$ and $f_1(x_1)$. Orthogonality ensures that the effects do not overlap and each term represents the isolated contribution.

Variance decomposition

The variance decomposition is Sobol's major use of fANOVA. He built the Sobol indices for sensitivity analysis on it. We sketch the variance decomposition here and note that it is only possible under independence assumption.
If $f \in \mathcal{L}^2$, then $f_{i_1,\ldots,i_n} \in \mathcal{L}^2$ 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:

$$\begin{aligned}
\sigma &= \int f^2(x)d\nu(x) - (f_0)^2 \\
&= \int f^2(x)d\nu(x) - (\int f(x)d\nu(x))^2 \\
&= \mathbb{E}[f^2(x)] - \mathbb{E}[f(x)]^2
\end{aligned}$$

The variance of the fANOVA components is then defined as

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

Because of the zero-mean constraint (Equation 2) the second term vanishes and we get

$$\begin{aligned}
\sigma(x_{i_1}, \ldots, x_{i_n}) &= \int \cdots \int f^2_{i_1,\ldots,i_n}\, d\nu(x_1) \cdots d\nu(x_n) \\
&= \mathbb{E}[f^2_{i_1,\ldots,i_n}]
\end{aligned}$$

With the definition of the total variance D and the component-wise variance $D_{i_1,\ldots,i_n}$ we can now see that the total variance can be decomposed into the sum of the component-wise variances.

We come back to our example $g(x_1, x_2)$ to illustrate the variance decomposition.

$$\begin{aligned}
\sigma &= \int g^2(x_1, x_2)\, d\nu(x) - f_0^2 \\
&= \mathbb{E}[g^2(x_1, x_2)] - a^2 \\
&= \mathbb{E}[(x_1 + 2x_2 + x_1x_2 + a)^2] - a^2 \\
&= \mathbb{E}\left[x_1^2 + 4x_2^2 + x_1^2x_2^2 + a^2 + 4x_1x_2 + 2x_1^2x_2 + 2ax_1 + 4x_1x_2^2 + 4ax_2 + 2ax_1x_2\right] - a^2 \\
&= \mathbb{E}[x_1^2] + 4\mathbb{E}[x_2^2] + \mathbb{E}[x_1^2x_2^2] + 4\mathbb{E}[x_1x_2] + 2\mathbb{E}[x_1^2x_2] + 2a\mathbb{E}[x_1] + 4\mathbb{E}[x_1x_2^2] + 4a\mathbb{E}[x_2] + 2a\mathbb{E}[x_1x_2] \\
&= \sigma^2(x_1) + 4\sigma^2(x_2) + \sigma^2(x_1x_2) + 2\mathbb{E}[x_1^2x_2] + 4\mathbb{E}[x_1x_2^2]
\end{aligned}$$

This holds because:

$$\begin{aligned}
\sigma(X_1) &= \mathbb{E}[X_1^2] - (\mathbb{E}(X_1))^2 = \mathbb{E}[X_1^2] \\
4\sigma(X_2) &= \sigma(2X_2) = \mathbb{E}[(2X_2)^2] - (\mathbb{E}(2X_2))^2 = \mathbb{E}[(2X_2)^2] \\
\sigma(X_1X_2) &= \mathbb{E}[X_1^2X_2^2] - (\mathbb{E}[X_1X_2])^2 = \mathbb{E}[X_1^2X_2^2]
\end{aligned}$$

Notice that we used the independence assumption and the zero-mean constraint again for the variance decomposition.

fANOVA as projection

Referring to the general connection between the expected value and orthogonal projections presented in section 2, the fANOVA terms can also be understood from a viewpoint of projections. This will also help to understand the generalization of fANOVA in section 4. f_0 is the projections of the original function f onto the space of all constant functions $\mathcal{G}_0 = \{g(x) = a; a \in \mathbb{R}\}$. It is an unconditional expected value and the best approximation of f given a constant function:

$$
\begin{aligned}
\Pi_{\mathcal{G}_0} f &= \arg\min_{g \in \mathcal{G}_0} \|f(x) - g\|^2 \\
&= \arg\min_{g \in \mathcal{G}_0} \mathbb{E}[\|f(x) - g\|^2] \\
&= \mathbb{E}[f(X)]
\end{aligned}
$$

The main effect $f_i(x_i)$ is the projection of f 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 $\mathcal{G}_i = \{g(x) = g_i(x_i); \int g(x) d\nu(x_i) = 0\}$.

$$
\begin{aligned}
\Pi_{\mathcal{G}_i} f - f_0 &= \arg\min_{g \in \mathcal{G}_i} \|f(x) - g(x_i)\|^2 - f_0 \\
&= \arg\min_{g \in \mathcal{G}_i} \mathbb{E}_{-x_i}[\|f(x) - g(x_i)\|^2] - \mathbb{E}[f(x)] \\
&= \mathbb{E}_{-x_i}[f(X_1, \dots, x_i, \dots, X_n)] - \mathbb{E}[f(X)]
\end{aligned}
$$

The two-way interaction effect $f_{ij}(x_i, x_j)$ is the projection of f 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:

$$
\begin{aligned}
\Pi_{\mathcal{G}_{ij}} f - f_0 - f_1(x_i) - \dots &= \arg\min_{g \in \mathcal{G}_{ij}} \|f(x) - g(x_i, x_j)\|^2 - f_0 - f_1(x_i) - \dots \\
&= \arg\min_{g \in \mathcal{G}_{ij}} \mathbb{E}_{-x_i, -x_j}[\|f(x) - g(x_i, x_j)\|^2] - \mathbb{E}[f(x)] - \mathbb{E}_{-x_i}[f(x)] \\
&= \mathbb{E}_{-x_i, -x_j}[f(X_1, \dots, x_i, x_j, \dots, X_n)] - \mathbb{E}[f(x)] - \mathbb{E}_{-x_i}[f(X)]
\end{aligned}
$$

I think Hilbert space theorem tells us that the orthogonal projection minimizes the squared difference in a Hilbert space? So the projection is the solution to the minimization problem that wants to minimize the squared differences between two elements of the vector space. This would be the first equality. The last equality that the solution is equal to the (conditional) expected value also has to be shown, still have to look which theorem this is proven by.

In general, general we can write:

$$f_u(x) = \Pi_{\mathcal{G}_u} f - \sum_{v \subsetneq u} f_v(x) \tag{13}$$

We project f onto the subspace spanned by the own terms of the fANOVA component to be defined, while accounting for all lower-order terms.

4 Generalization

The chapter is based on Hooker (2007). 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 1). What changes is the definition of the fANOVA components. The components are simultaneously defined as:

$$\{f_u(x_u) \,|\, u \subseteq d\} = \arg \min_{\{g_u \in L^2(\mathbb{R}^u)\}_{u \subseteq d}} \int \left(\sum_{u \subseteq d} g_u(x_u) - f(x) \right)^2 w(x)\, dx \tag{14}$$

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 hierarchical 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:

$$\forall u \subseteq d,\ \forall i \in u : \quad \int f_u(x_u)\, w(x)\, dx_i\, dx_{-u} = 0 \tag{15}$$

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):

$$\forall v \subseteq u,\ \forall g : \quad \int f_u(x_u)\ g_v(x_v) w(x)\ dx = 0 \tag{16}$$

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:

$$\forall v \subset u : \quad \int f_u(x_u)\ g_v(x_v) w(x)\ dx = 0 \tag{17}$$

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.

Example

Can I even calculate generalized fANOVA by hand? Because we need to optimize simultaneously? After introducing a framework for generalized fANOVA it is interesting to come back to our previous example $g(x_1, x_2) = a + x_1 + 2x_2 + x_1 x_2$ and see how the general decomposition works differently.

5 Simulation Study

5.1 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

6 Conclusion

7 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

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

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*

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*

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*

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