

Bachelor's Thesis

---

# fANOVA for Interpretable Machine Learning

---

Department of Statistics  
Ludwig-Maximilians-Universität München

**Juliet Fleischer**

Munich, Month Day<sup>th</sup>, 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

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Foundations</b>	<b>2</b>
2.0.1	Prerequisites . . . . .	4
2.0.2	fANOVA decomposition . . . . .	5
2.0.3	Example fANOVA decomposition . . . . .	8
<b>3</b>	<b>Conclusion</b>	<b>10</b>
<b>A</b>	<b>Appendix</b>	<b>V</b>
<b>B</b>	<b>Electronic appendix</b>	<b>VI</b>

# 1 Introduction

## 2 Foundations

### 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.
- Assumptions: unclear about all but one assumption is (mutual?) 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 Models" (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 estimates" (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 decomposition, 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.  $\Rightarrow$

textbfANOVA 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.

### **$\Rightarrow$ 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.

## Modern Interpretations of fANOVA

- Rabitz and Alis, (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 select 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.

## Formal Introduction to fANOVA

### 2.0.1 Prerequisites

We are operating the space of square integrable functions. The space of square integrable functions is denoted by  $\mathcal{L}^2$  and is defined as follows:

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

- $\mathbb{E}$  denotes the expectation operator
- $\mathbb{R}^n$  is the n-dimensional Euclidean space
- $\nu$  is the measure of the Lebesgue integral

For the following we will restrict ourselves to functions defined on the unit hypercube  $[0, 1]^n$ , i.e. all functions  $f(x) : [0, 1]^n \rightarrow \mathbb{R}$ , where  $[0, 1]^n \subset \mathcal{L}^2$   
 $\mathcal{L}^2$  is a Hilbert space with an inner product defined as

$$\langle f, g \rangle = \int f(x)g(x) d\mu(x)$$

where  $f, g \in \mathcal{L}^2$  are two functions.

The norm is then defined as

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int f^2(x) d\mu(x)}$$

- inputs  $x_1, \dots, x_n$  are independent
- we assume that each input is uniformly distributed over the unit interval, i.e.  $x_i \sim U[0, 1]$  for  $i = 1, \dots, n$  (but this is not necessary, independence is though (ANOVA lecture notes by Owen))

could cite this resource for general definition of Normed vectors paces, Hilbert space, inner product, etc. <https://apachepersonal.mium.se/~andrli/Bok.pdf?>

### 2.0.2 fANOVA decomposition

This chapter is based on the formal introductions by Hoeffding (1948), Sobol (2001), Hooker (2004). Let  $f(x)$  be a mathematical model.  $f(x) \in \mathcal{L}^2$ , which means  $\int |f(x)|^2 < \infty$ .  $f(x)$  is a multivariate function with input  $x = (x_1, \dots, x_n)$  and output  $f(x) \in \mathbb{R}$ . In order for fANOVA decomposition to be unique, the function has to be in the unit hypercube, i.e.  $f : [0, 1]^n \rightarrow \mathbb{R}$  (reference?).

**Definition.** We can represent such a function  $f(x)$  as a sum of specific basis functions

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

To ensure identifiability and interpretation, we set the zero-mean constraint. It requires that all effects are centred around zero<sup>1</sup>. Since that the constant term  $f_0$  captures the overall mean of  $f$ , the remaining effects quantify the deviation from the overall mean.

$$\int_0^1 f_{i_1 \dots i_s}(x_{i_1}, \dots, x_{i_s}) dx_k = 0 \forall k = i_1, \dots, i_s \quad (2)$$

In combination with Equation 2 I.M. 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 basis components as they are defined offer a clear interpretation of the model. They are nothing other than the main effects  $f_i(x_i)$ , two-way interaction effects  $f_{ij}(x_i, x_j)$ ,

---

<sup>1</sup>Except for the constant term



three-way interaction effects  $f_{ijk}(x_i, x_j, x_k)$ , and so on. This is why the idea of fANOVA decomposition has received increasing attention in the IML and XAI literature.

The single terms that make up Equation 1 are defined as follows. First, we take the integral of  $f$  w.r.t. all variables:

$$f_0(\mathbf{x}) = \int_{[0,1]^n} f(\mathbf{x}) d\nu(\mathbf{x}) = \mathbb{E}[f(x)] \quad (3)$$

Next, we take the integral of  $f$  w.r.t. all variables except for  $x_i$ . This gives us the sum of the constant term and the isolated effect of  $x_i$ , called main effect of  $x_i$ .

$$f_0 + f_i(x_i) = \int f(x) \prod_{k \neq i} d_{x_k} = g_i(x_i) \quad (4)$$

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  $x_j$ :

$$f_0 + f_i(x_i) + f_j(x_j) + f_{ij}(x_i, x_j) = \int f(x) \prod_{k \neq i, j} d_{x_k} = g_{ij}(x_i, x_j) \quad (5)$$

And so on, up to the n-way interaction of all  $x_1, \dots, x_n$ .

To actually compute the fANOVA decomposition for  $f$ , it is clearer to rearrange terms.

We write Equation 4 as

$$f_i(x_i) = \int f(x) \prod_{k \neq i} d_{x_k} - f_0 \quad (6)$$

Then Equation 7 becomes

$$f_{ij}(x_i, x_j) = \int f(x) \prod_{k \neq i, j} d_{x_k} - f_0 - f_i(x_i) - f_j(x_j) \quad (7)$$

Intuitively we adjust for the effects already explained by lower order terms by subtracting them from [the overall effects \(not true, not the overall but a partial effect, a conditional effect\)](#).

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

$$f_u(\mathbf{x}) = \int_{[0,1]^{d-|u|}} \left( f(\mathbf{x}) - \sum_{v \subsetneq u} f_v(\mathbf{x}) \right) d\nu(\mathbf{x}_{-u}). \quad (8)$$

Which simplifies to:

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

A technical remark: The fANOVA terms can be understood as projections.

$f_0$  is the projections of  $f$  onto the space of all constant functions  $G = \{g(x) = a; a \in \mathbb{R}\}$ . It is an unconditional expected value and the best approximation of  $f$  given by a constant function.

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,  $G = \{g(x) = g_i(x_i); \int g(x) d\nu(x_i) = 0\}$ . It is a mean conditioned on  $x_i$  and the best approximation of  $f$  given by a function that depends on a single variable  $x_i$ .

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 its single components,  $G = \{g(x) = g_{ij}(x_i, x_j); \int g(x) d\nu(x_i) = 0 \wedge \int g(x) d\nu(x_j) = 0\}$ . It is the expected value conditioned on  $x_i, x_j$  and the best approximation of  $f$  given by a function that depends on only two variables.

In general, “each term is calculated as the projection of  $f$  onto a particular subset of the predictors, taking out the lower-order effects which have already been accounted for.” Hooker (2004).

### Orthogonality of the fANOVA terms

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

$$\int f_{i_1, \dots, i_s} f_{j_1, \dots, j_l} dx = 0 \quad (10)$$

### Variance decomposition

If  $f \in \mathcal{L}^2$ , it follows that  $f_{i_1, \dots, i_n} \in \mathcal{L}^2$  [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:

$$D = \int_{K^n} f^2(x) d\nu(x) - f_0^2 = \int_{K^n} f^2(x) d\nu(x) - \left( \int_{K^n} f(x) d\nu(x) \right)^2 \quad (11)$$

The variance of the fANOVA components is then defined as

$$\begin{aligned} D_{i_1, \dots, i_n} &= \int \cdots \int f_{i_1, \dots, i_n}^2 d\nu(x_1) \cdots d\nu(x_n) \\ &\quad - \left( \int \cdots \int f_{i_1, \dots, i_n} d\nu(x_1) \cdots d\nu(x_n) \right)^2 \\ &= \int \cdots \int f_{i_1, \dots, i_n}^2 d\nu(x_1) \cdots d\nu(x_n) \end{aligned}$$

This means that the variance of  $f$  can be decomposed into the variance of the fANOVA components. Sobol 1993 uses this result to construct the Sobol Indices used in sensitivity analysis.

### 2.0.3 Example fANOVA decomposition

We will now make the fANOVA decomposition of a function  $f \in \mathcal{L}^2$  explicit for  $n = 4$ . We continue to follow the definition by Sobol (2001).

## Questions

- Use of AI tools?
- —————
- 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.*

## Notes

- decomposition always exists
- zero-mean-condition  $\rightarrow$  orthogonality of the terms
- $K^n$ -integrable functions  $\rightarrow$  uniqueness of the decomposition
- when does the variance decomposition exist? I think this is related to square integrability, i.e.  $L^2$  integrable<sup>2</sup>, which means that the integral of the square of the function is finite
- keep in mind: projections, hierarchical orthogonality constraints

---

<sup>2</sup> $L^1$  integrable does not imply  $L^2$  integrable, and vice versa

### 3 Conclusion

# 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.
- 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>
- 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>
- 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