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 **1**

2 Foundations **2**

2.0.1 Prerequisites 4

2.0.2 fANOVA decomposition 5

2.0.3 Example fANOVA decomposition 9

3 Conclusion **11**

A Appendix **V**

B Electronic appendix **VI**

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.
- 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 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 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. ⇒
textbffANOVA 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.

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 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.

Formal Introduction to fANOVA

2.0.1 Prerequisites

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

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

- $\mathbb{E}$ denotes the expectation operator
- $\mathbb{R}^n$ is the n-dimensional Euclidean space
- ν 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 $f(x) : [0,1]^n \to \mathbb{R}$, while $[0,1]^n \subset \mathcal{L}^2$. $\mathcal{L}^2$ is a Hilbert space with an inner product defined as

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

The norm is then defined as

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

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

2.0.2 fANOVA decomposition

This chapter is based on the formal introductions by Hoeffding (1948), Sobol (1993, 2001), Hooker (2004). Let $f : [0,1]^n \to \mathbb{R}$ be a mathematical model. $f \in \mathcal{L}^2$, which means $\int |f(x)|^2 < \infty$. We assume for the input $x = (x_1, \dots, x_n)$ that the x_i are independent from each other and $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$.

Definition. We can represent such a model f as a sum of specific basis functions

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

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. Since that the constant term f_0 captures the overall mean of f, the remaining effects quantify the deviation from the overall mean. Mathematically this means

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

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

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

Next, we take the integral of f w.r.t. all variables except for x_i. This represents f as the sum of the constant term the isolated effect of one variable x_i (main effect of x_i).

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

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:

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

And so on, up to the n-way interaction of all $x_1, \ldots, x_n$.
To actually compute the fANOVA decomposition for f, it is clearer to rearrange terms. When we rearrange Equation 4 we get that the main effect of x_i is calculated by taking the marginal effect while explicitly accounting for what was already explained by lower terms, in this case the intercept.

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

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.

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

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

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

Which simplifies to:

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

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.
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 my 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 it's 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 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

$$\int f_{i_1,\ldots,i_s} f_{j_1,\ldots,j_l} dx = 0 \tag{10}$$

This means that fANOVA terms are "fully orthogonal" to each other.
Consider an example for $(i_1, i_2) = (1, 2)$ and $(j_1, j_2) = (1, 3)$. We take the inner product between these fANOVA components

$$\int_0^1 \int_0^1 \int_0^1 f_{1,2}(x_1, x_2) \cdot f_{1,3}(x_1, x_3)\, dx_1\, dx_2\, dx_3$$

We begin by integrating with respect to x_1 and define

$$\int_0^1 f_{1,2}(x_1, x_2) \cdot f_{1,3}(x_1, x_3)\, dx_1 := h(x_2, x_3)$$

Then the full integral becomes

$$\int_0^1 \int_0^1 \left(\int_0^1 f_{1,2}(x_1, x_2) \cdot f_{1,3}(x_1, x_3)\, dx_1 \right) dx_2 dx_3 = \int_0^1 \int_0^1 h(x_2, x_3)\, dx_2\, dx_3$$

But can I now simply say that the integral wrt to x_2 is zero because of the zero-mean constraint Equation 2?

$$\int_0^1 h(x_2, x_3)\, dx_2 = 0 \quad \text{for all } x_3$$

Variance decomposition

If $f \in \mathcal{L}^2$, then $f_{i_1,\dots,i_n} \in \mathcal{L}^2$ 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:

$$D = \int_{K^n} f^2(x)d\nu(x) - f_0^2 = \int_{K^n} f^2(x)d\nu(x) - (\int_{K^n} f(x)d\nu(x))^2 \quad = \mathbb{E}[f^2(x)] - \mathbb{E}[f(x)]^2$$

The variance of the fANOVA components is then defined as

$$D_{i_1,\dots,i_n} = \int \cdots \int f^2_{i_1,\dots,i_n}\, d\nu(x_1)\cdots d\nu(x_n) - \left(\int \cdots \int f_{i_1,\dots,i_n}\, d\nu(x_1)\cdots d\nu(x_n)\right)^2$$

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

$$D_{i_1,\dots,i_n} = \int \cdots \int f^2_{i_1,\dots,i_n}\, d\nu(x_1)\cdots d\nu(x_n)$$

With the definition of the total variance D and the component-wise variance $D_{i_1,\dots,i_n}$ we can now see that the total variance can be decomposed into the sum of the component-wise variances.
We illustrate this for a fANOVA decomposition function $f(x_1, x_2) \in L^2$:

$$f(x_1, x_2) = f_0 + f_1(x_1) + f_2(x_2) + f_{1,2}(x_1, x_2)$$

where:

- $f_0 \in \mathbb{R}$ is a constant,
- $f_1(x_1)$, $f_2(x_2)$, and $f_{1,2}(x_1, x_2)$ have zero mean in their own variables,
- all components are mutually orthogonal in $L^2([0, 1]^2)$.

Step 1: Square the decomposition.

$$\begin{aligned} f^2(x_1, x_2) &= (f_0 + f_1(x_1) + f_2(x_2) + f_{1,2}(x_1, x_2))^2 \\ &= f_0^2 + f_1(x_1)^2 + f_2(x_2)^2 + f_{1,2}(x_1, x_2)^2 \\ &\quad + 2f_0f_1(x_1) + 2f_0f_2(x_2) + 2f_0f_{1,2}(x_1, x_2) \\ &\quad + 2f_1(x_1)f_2(x_2) + 2f_1(x_1)f_{1,2}(x_1, x_2) + 2f_2(x_2)f_{1,2}(x_1, x_2) \end{aligned}$$

Step 2: Integrate over the domain $[0,1]^2$.

$$\begin{aligned}
\int f^2(x_1,x_2)\,dx_1dx_2 &= \int f_0^2\,dx_1dx_2 + \int f_1(x_1)^2\,dx_1dx_2 + \int f_2(x_2)^2\,dx_1dx_2 \\
&\quad + \int f_{1,2}(x_1,x_2)^2\,dx_1dx_2 \\
&\quad + \text{(all cross-terms vanish due to orthogonality)} \\
&= f_0^2 + \int f_1(x_1)^2\,dx_1 + \int f_2(x_2)^2\,dx_2 + \int f_{1,2}(x_1,x_2)^2\,dx_1dx_2
\end{aligned}$$

Step 3: Define the variance.

$$D := \int f(x_1,x_2)^2\,dx_1dx_2 - f_0^2$$

Then the total variance decomposes as:

$$D = D_1 + D_2 + D_{1,2}$$

where:

$$D_1 = \int f_1(x_1)^2\,dx_1, \quad D_2 = \int f_2(x_2)^2\,dx_2, \quad D_{1,2} = \int f_{1,2}(x_1,x_2)^2\,dx_1dx_2$$

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[1], which means that the integral of the square of the function is finite

- keep in mind: projections, hierarchical orthogonality constraints

[1] 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*

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