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

2 Foundations

Early work on fANOVA

- 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

- 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.
- 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 1) ranking of variables; 2) fixing unessential variables; 3) deleting high order members; for each of the three there are some mathematical statements, sometimes an algorithm or an example. In I.M. Sobol (1993) Theorem 1 states that for any function $f(x)$ which is integrable on K^n, there exists a unique expansion of Equation 1 (proof follows after the theorem in the same paper).

fANOVA and sensitivity analysis

- 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 Setting of fANOVA

- we look at a function (which represents a mathematical model) from high dimensional space (or often the high dimensional unit hypercube, without loss of generality) to reel number line
- in Hooker (2004, 2007) the function is square integrable/ is L^2
- I think in the basic setting of Sobol (2001) the function is not square integrable per se but if it is then desirable properties are met or sth.
- Lecture Notes on Sensitivity Analysis already say that the input X is uncorrelated

Let $f(x) : I[0,1]^n \to I[0,1] \quad with \quad x = (x_1, ...x_n)$ be a function from the unit hypercube to the unit interval. $f(x)$ represents a mathematical model.
Definition: $f(x)$ can be represented as a sum of main effects and interaction effects

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

with $1 \leq i_1 < .... < i_s \leq n$. Equation 1 is the "ANOVA-representation" Sobol (2001) or functional ANOVA decomposition Hooker (2004) is f_0 is constant and the integrals of the summands $f_{i_1 ... i_s}$ with respect to any of their included variables are zero.

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

In words so far: a function is decomposed into constant term f_0 and a sum of main effects and interaction effects. If each term "is centred" (i.e. has zero mean) with respect to the variables it includes, then the terms are orthogonal to each other. In an applied context orthogonality means that the terms capture the isolated effect and there is no redundancy in information, i.e. no information of x_1 is also included in the interaction of $x_1 2$ "Expansion into summands of different dimensions" (I.M. Sobol, 1993)/ "fANOVA decomposition"(Hooker, 2004) **Example for** $n = 3$:

$$f(x_1, x_2, x_3) = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) + f_{13}(x_1, x_3) + f_{23}(x_2, x_3) \tag{3}$$

We can use Equation 2 and Equation 1 to formulate expressions for specific components of $f(x)$, i.e.:

$$\int f(x) dx = f_0 \tag{4}$$

This means that the integral over the entire domain and all inputs gives us the constant term/ intercept == overall average

$$\int f(x) \prod_{k \neq i} d_{x_k} = f_0 + f_i(x_i) \tag{5}$$

The integral over all variables expect x_i is equal to adding up the overall mean and the main effect of x_i.

$$\int f(x) \prod_{k \neq i,j} d_{x_k} = f_0 + f_i(x_i) + f_j(x_j) \tag{6}$$

The integral over all variables expect x_i and x_j is equal to adding up the overall mean, the main effect of x_i, and the main effect of x_j, and the interaction effect of x_i, x_j.
Inhaltlich: when we are interested in "the average effect of ..." we may add the main effect and their interaction effects together.??

An attempt to explain the motivation behind fANOVA

Disclaimer: Intuitive explanation.

To motivate the fANOVA decomposition, we tell the story "backwards"[1]. We start with a statistical model $f(x)$ that takes in multiple covariables $x_1, \ldots, x_n$ that are potentially interacting with each other in a complex way and having a complex effect on the target variable. We therefore think, it would be very nice to disentangle the effects of the covariables and clearly tell which influence comes from a single variable, which comes from interactions and so on. Mathematically this would correspond to writing $f(x)$ as a sum of isolated terms and all n-way interaction terms, i.e. Equation 1.
The challenge becomes to find this specific representation of $f(x)$. In this sense we are facing an approximation problem, which we solve by using projections (we need the best approximation of $f(x)$ in a certain subspace, so the orthogonal projection onto the subspace is the best solution). A projection is defined via the inner product, which is defined via integrals in case of functions. The projections we use to define the components of the decomposition therefore look like this:

$$\int f(x)dx = f_0 \tag{7}$$

This means that the integral over the entire domain and all inputs gives us the constant term/ intercept == overall average

$$\int f(x) \prod_{k \neq i} d_{x_k} = f_0 + f_i(x_i) \tag{8}$$

The integral over all variables expect x_i is equal to adding up the overall mean and the main effect of x_i.

$$\int f(x) \prod_{k \neq i,j} d_{x_k} = f_0 + f_i(x_i) + f_j(x_j) \tag{9}$$

With the tool of projections we solved the approximation problem. Next we face the problem of ensuring this approximation exists. We know that a representation with orthogonal projections always exists for function in L^2. Thus, we choose to restrict the functions we look at to be in L^2. This brings us to define the fANOVA decomposition as follows.

An attempt to formalize fANOVA

Let $f(x)$ be a mathematical model. $f(x) \in L^2$, which means $\int |f(x)|^2 < \infty$. $f(x)$ is a multivariate function with input $x = (x_1, \ldots, x_n)$ and output $f(x) \in R$. We can represent $f(x)$ as a sum of specific orthogonal basis functions:Equation 1.
The basis components are constructed via projections:

$$\int f(x)dx = f_0 \tag{10}$$

This means that the integral over the entire domain and all inputs gives us the constant term/ intercept == overall average

$$\int f(x) \prod_{k \neq i} d_{x_k} = f_0 + f_i(x_i) \tag{11}$$

[1] In the opposite way, it is usually introduced formally.

The integral over all variables expect x_i is equal to adding up the overall mean and the main effect of x_i.

$$\int f(x) \prod_{k \neq i,j} d_{x_k} = f_0 + f_i(x_i) + f_j(x_j) \tag{12}$$

We set the zero-mean constraint to ensure that the basis components are orthogonal, i.e. Equation 2. 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 $f_{ij}(x_i, x_j)$, three-way interaction effects $f_{ijk}(x_i, x_j, x_k)$, and so on.

- decomposition always exists
- zero-mean-condition → orthogonality of the terms
- K^n-integrable functions → uniqueness of the decomposition (L^1 integrable, which means that the integral of the absolute value of the function is finite)
- when does the variance decomposition exist? I think this i related to square integrability, i.e. L^2 integrable[2], which means that the integral of the square of the function is finite
- keep in mind: projections, hierarchical orthogonality constraints
- orthogonal projections
- function spaces, finite-dimensional spaces
- in Hooker (2004) they work with $F(x)$ and $f(x)$, but in Sobol (2001) they only work with $f(x)$
- zero mean condition vs. zero-sum condition: according to GPT zero mean condition is related to orthogonality and zero-sum to the additivity
- does orthogonality mean that all terms are orthogonal to each other? or that a term is orthogonal to all lower-order terms?
- function space vs. finite vector space, projections in both of these, projections as integrals
- is this related to projections as approximations (linalg skript 2024 5.9.4 Funktionenräume)
- in fANOVA decomposition, do the fANOVA members constitute an orthogonal basis for the function space in which the model function lives? In which space does the model function live - L2? Could it be that we construct the orthogonal basis "artificially" with projections?

[2] 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

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