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.
- 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.

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- All of them work independently, do not cite each other, and use the principle with different goals/build different tools on it.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

Modern Interpretations of fANOVA

- 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 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. (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