Fairness seen as Global Sensitivity Analysis

C. Bénesse
Université Paul Sabatier, Toulouse, France

Supervisor(s): Prof. F. GAMBOA (Université Paul Sabatier, Toulouse), Prof. J-M. LOUBES (Université Paul Sabatier, Toulouse)

Ph.D. expected duration: Oct. 2019 - Sep. 2022

Adress: Université Paul Sabatier, Institut de Mathématiques de Toulouse, 118 route de Narbonne - F-31062 TOULOUSE Cedex 9

Email: clement.benesse@math.univ-toulouse.fr

Abstract:

Quantifying the influence of a variable on the outcome of an algorithm is an issue of high importance in order to understand decisions and detect unwanted biases in the decisions that may lead to unfair predictions. The recent development of the literature on fair learning for Artificial Intelligence shows how essential this problem is. One of the main difficulty lies in the definition of what is (un)fair and the choices to quantify it. Historically, Statistical Parity has been used to quantify fairness in legal procedures. This measure has then been used in machine learning ([2]) to assess the fairness of an estimator's output. Similarly, Equality of odds can be used to quantify fairness by looking at the errors made by the estimator. However, these measures restrict the notion of fairness between two discrete classes in the context of a classification problem. Authors of [5], [9] have extended these notions to regression problems, but also with continuous sensitive attributes. Independently other metrics such as Shapley values has been used in the context of fairness [6] (local fairness by explainability).

Conversely, Global Sensitivity Analysis (GSA) is used in numerous contexts for quantifying the influence of a set of features on the outcome of a black-box algorithm. Various indicators, usually taking the form of indices between 0 and 1, allow the understanding of how much a feature is important. Multiple set of indices have been proposed over the years such as Sobol' indices, Cramér-von-Mises indices, HSIC – see [1], [3], [7], [8], [4] and references therein. The flexibility in the choice allows for deep understanding in the relationship between a feature and the outcome of an algorithm. While a usual assumption in this field is to suppose the inputs to be independent, some works ([10], [4], [8]) remove this assumption to go further in the understanding of the possible ways for a feature to be influent.

Our contributions are two-fold. Firstly, while GSA is usually concerned with independent inputs, we recall extensions of Sobol' indices to non-independent inputs introduced in [10] that offer ways to account for joint contribution and correlations between variables while quantifying the influence of a feature. We propose an extension of Cramér-von-Mises indices based on similar ideas. We also prove the asymptotic normality for these extended Sobol' indices. Secondly, thanks to Global Sensitivity Analysis, it is possible to monitor the importance of a feature on the outcome of an algorithm. Therefore, by using it in a Fairness framework, we are able to quantify how fair an algorithm is with respect to a sensible feature. We will see that different sets of GSA indices lead to different definitions of Fairness and we propose a framework for quantifying intersectionality fairness. Thanks to this duality, this link allows for a better understanding of the relationship between a protected feature and the outcome of a predictor.

References

[1] Sébastien Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, May 2015.

[2] Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

[3] Fabrice Gamboa, Pierre Gremaud, Thierry Klein, and Agnès Lagnoux. Global sensitivity analysis: a new generation of mighty estimators based on rank statistics. *arXiv preprint arXiv:2003.01772*, 2020.

[4] Mathilde Grandjacques. *Analyse de sensibilité pour des modèles stochastiques à entrées dépendantes: application en énergétique du bâtiment.* PhD thesis, Grenoble Alpes, 2015.

[5] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural réyni minimization for continuous features, 2019.

[6] James M. Hickey, Pietro G. Di Stefano, and Vlasios Vasileiou. Fairness by explicability and adversarial shap learning, 2020.

[7] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015.

[8] Julien Jacques, Christian Lavergne, and Nicolas Devictor. Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering & System Safety*, 91(10-11):1126–1134, 2006.

[9] Noureddine El Karoui Jeremie Mary, Clement Calauzenes. Fairness-aware learning for continuous attributes and treatments, 2019.

[10] Thierry A Mara and Stefano Tarantola. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety*, 107:115–121, 2012.

Short biography – I am a Ph.D Student at *Institut de Mathématiques de Toulouse* and *ANITI*, under the supervition of *Fabrice GAMBOA* and *Jean-Michel LOUBES*. I graduated from *École Normale Supérieure de Lyon* in 2019. I am currently working on Global Sensitivity Analysis (especially Sobol' and Cramér-von-Mises indices), social and industrial Fairness (including intersectionality), and the links between these fields.