Efficient estimation of the Shapley sensitivity indices for the linear Gaussian model with independent groups of variables.

Baptiste Broto
CEA Saclay

Supervisor(s): Marine Depecker (CEA), François Bachoc (IMT) and Jean-Marc-Martinez (CEA)

Ph.D. expected duration: sept. 2017 - sept. 2020

Adress: CEA Saclay, 91191, Gif-sur-Yvette CEDEX

Email: baptiste.broto@cea.fr

Abstract: In this research work, we focus on the computation of the Shapley values for global sensitivity analysis of Gaussian linear models with independent groups of variables. The main contribution of this work relies on a proposition that considerably reduces the computational time in this specific context. We also propose numerical experiments to illustrate this theoretical contribution. In particular, we suggest an algorithm exploiting the proved proposition and we compare it with state-of-art algorithms for the computation of Shapley values. Theses experiments highlight the specificity of our approach, which is the only one providing an exact computation of the Shapley values in high dimension.

Recently, Owen used the notion of Shapley value (which originates from game theory) in order to define new sensitivity indices (see [4]) preserving good properties when the input variables are dependent, in particular they are always in $[0, 1]$. Here, we focus on computing the Shapley values in the linear Gaussian case, a framework which is widely used to model physical phenomena (see e.g. [3]). As the Shapley values are based on the conditional variances, they can be computed explicitly in this framework (see [5] and [2]):

$$\eta_i := \frac{1}{p\mathrm{Var}(Y)} \sum_{u \subset [1:p] \setminus \{i\}} \binom{p-1}{|u|}^{-1} \left(\mathrm{Var}(Y|X_u) - \mathrm{Var}(Y|X_{u\cup\{i\}})\right). \tag{1}$$

We introduce a first algorithm (Algorithm 1) to compute the Shapley values in this framework. However, because of Equation (1), the computation increases exponentially with the number of inputs p: each Shapley value requires the computation of 2^p conditional variances. Fortunately, when p is large, it can frequently be the case that there are independent groups of random variables. We thus propose to exploit the independence structure to reduce the computational cost of the Shapley values. We show that the high dimensional computational problem then boils down to a collection of lower dimensional problems.

Proposition 1 *Let $C_1, ..., C_k$ be a partition of $[1 : p]$ so that the random variables $X_{C_1}, ..., X_{C_k}$ are independent between them. Let $i \in [1 : p]$. Let $j(i) \in \mathbb{N}$ be so that $i \in C_{j(i)}$. Then, we have:*

$$\eta_i = \frac{1}{|C_{j(i)}|\mathrm{Var}(Y)} \sum_{u \subset C_{j(i)} \setminus \{i\}} \binom{|C_{j(i)}|-1}{|u|}^{-1} \left(\mathrm{Var}(Y|X_u) - \mathrm{Var}(Y|X_{u\cup\{i\}})\right). \tag{2}$$

The idea is that, the Shapley value of a variable X_i will only depend on the variables belonging to the same group as X_i. Hence, the computation of all Shapley values requires $\sum_{j=1}^{k} 2^{|C_k|}$

conditional variances instead of 2^p. We thus derive a second algorithm (Algorithm 2) that exploits this configuration.

To position our work with respect to the state of art, we compare Algorithm 1 and 2 with existing algorithms designed to compute the Shapley values for global sensitivity analysis. In particular, we consider the two functions "shapleyPermEx" and "shapleyPermRand" in the R package "sensitivity" described in [6]. We adapt these algorithms to the linear Gaussian framework for a fair comparison: the estimations of the conditional variances are replaced by their explicit formula. The first function "shapleyPermEx" computes theoretical values but the computational time can be important, while the second function "ShapleyPermRand" provides estimates with a limited computational time. The interest of Algorithm 2 is that it can compute the theoretical Shapley values with a limited computational cost.
In our first comparison, we show that in the general linear Gaussian framework, Algorithm 1 is already faster than "shapleyPermEx". Then, we compare Algorithm 2 with Algorithm 1 and "shapleyPermEx" for independent groups of variables, and we can see that Algorithm 2 is much faster. These results highlight the impact of Proposition 1 in the computation of Shapley values. We also perform numerical experiments in high dimension (up to a few hundred of inputs), so as to assess the behaviour of the different algorithms. In that context, Algorithm 1 and "shapleyPermEx" have a very high computational cost (several hours for only 25 inputs). In the meantime, Algorithm 2 provides the theoretical Shapley values in a few seconds at most, depending on the number of inputs and of independent groups. For the same computational time, the updated R function "shapleyPermRand" computes estimates with quite large coefficients of variation.

This work has been applied in the industrial field of nuclear safety, the achieved results have been presented by Pietro Mosca during the sixteen International Symposium on Reactor Dosimetry and submitted to the PHYSOR 2018 conference (see [1]).

References

[1] L. Clouvel, P. Mosca, and J.M. Martinez. Uncertainty propagation of double-differential scattering cross section in fast fluence calculation for the pwr surveillance capsules. In *submitted to PHYSOR 2018: Reactor Physics paving the way towards more efficient systems*, 2018.

[2] Bertrand Iooss and Clémentine Prieur. Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol' indices, numerical estimation and applications. *arXiv:1707.01334 [math, stat]*, July 2017. arXiv: 1707.01334.

[3] T. Kawano, K. M. Hanson, S. Frankle, P. Talou, M. B. Chadwick, and R. C. Little. Evaluation and Propagation of the ^{239}Pu Fission Cross-Section Uncertainties Using a Monte Carlo Technique. *Nuclear Science and Engineering*, 153(1):1–7, May 2006.

[4] A. Owen. Sobol' Indices and Shapley Value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, January 2014.

[5] Art B. Owen and Clémentine Prieur. On Shapley value for measuring importance of dependent inputs. *arXiv:1610.02080 [math, stat]*, October 2016.

[6] E. Song, B. Nelson, and J. Staum. Shapley Effects for Global Sensitivity Analysis: Theory and Computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, January 2016.

Short biography – I am PhD student in the LADIS laboratory of CEA LIST. I studied pure mathematics in the Université Paul Sabatier in Toulouse before to specialize in applied mathematics.