New estimation of sensitivity indices using kernels

Agnès Lagnoux[1]

[1]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT2J, F-31058 Toulouse, France.

The use of complex computer models for the simulation and analysis of natural systems from physics, engineering and other fields is by now routine. These models usually depend on many input variables.Thus, it is crucial to understand which input parameter or which set of input parameters have an influence on the output. This is the aim of sensitivity analysis which has become an essential tool for system modeling and policy support. Global sensitivity analysis (GSA) methods considers the input vector as random and propose a measure of the influence of each subset of its components on the output of interest. We refer to the seminal book by Saltelli [13] for an overview on GSA or to [4] for a synthesis of recent trends in the field. Among the different measures of GSA, variance-based measures are probably the most commonly used. The definition of the so-called Sobol' indices, introduced in [12] and later revisited in the framework of sensitivity analysis in [14], is based on the Hoeffding decomposition of the variance [10]. More precisely, for $Y = f(X_1, \ldots, X_p)$ where the inputs X_i are assumed to be mutually independent, the Sobol' index with respect to X is defined by

$$S^X = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} = \frac{\mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2}{\mathrm{Var}(Y)}$$

where X is a subvector of $(X_1, \ldots, X_p)$.
In general, computing explicitly the theoretical value of S^X is hopeless; therefore one of the main tasks is to provide estimators of S^X that have some nice convergence properties (consistency, explicit rate of convergence...). Many different estimation procedures were developped in the past decades to estimate S^X. Two classes of methods have emerged with nice theoretical convergence guarantees.

The first class is based on the so-called Pick Freeze (PF) design of experiments, the theoretical properties of which (consistency, central limit theorem, concentration inequalities and Berry-Esseen bounds) have been studied in [8, 11]. This method has very nice convergence properties and is very general since the only assumption needed to prove a central limit theorem is that $\mathbb{E}[Y^4] < \infty$. However, it requires the evaluation of the model on a PF design that may be unvailable in practice. Also, the required sample size increases linearly with the number of indices that one wants to estimate.

The second class of methods is based on nearest neighbors. In [6, 5], the authors built an estimator of $\mathbb{E}[\mathbb{E}[Y|X]^2]$ based on two independent n-samples. The first one allows to estimate the regression function $\mathbb{E}[Y|X = x]$ using the first nearest neighbor of x among the first sample and the second one is used as a plug-in estimator. They proved a central limit theorem for their estimator for $d \leqslant 3$. In [2], the author introduced a new coefficient of correlation based on ranks for $d = 1$. This approach was then adapted for building estimators of first-order Sobol' indices in [7], with a central limit theorem in the framework $d = 1$ and X_i not necessarily independent from Y. Broto *et al.* [1] also considered a nearest neighbor approach to estimate Sobol' indices of any order. The estimators they introduced are consistent but no rate of convergence was provided. Although very powerfull it is clear thanks to the bias study in [5] that it is not possible to obtain a central limit theorem for estimating Sobol' indices of order larger than $d = 3$ using this approach. It is important to note that all the procedures based on nearest neighbors require additional regularity assumptions to control the bias.

Close to nearest neighbor approaches are kernel methods. A natural estimator is the Nadaraya-Watson plug-in estimator. The analysis of such estimators also require regularity assumptions on the model. In this work, we develop an approach mentioned in [3, Remark 1]. It consists in extending the very interesting point of view introduced in [9] to estimate general nonlinear integral functionals of a density on the real line, by using empirically a kernel estimator erasing the diagonal terms. Relaxing the positiveness assumption on the kernel and choosing a kernel of order large enough, we are able to prove a central limit theorem for estimating Sobol' indices of any order (the bias is killed thanks to this signed kernel).

References

[1] B. Broto, F. Bachoc, and M. Depecker. Variance reduction for estimation of shapley effects and adaptation to unknown input distribution. SIAM/ASA Journal on Uncertainty Quantification, 8(2):693–716, 2020.

[2] S. Chatterjee. A new coefficient of correlation. Journal of the American Statistical Association, pages 1–26, 2020.

[3] S. Da Veiga and F. Gamboa. Efficient estimation of sensitivity indices. Journal of Nonparametric Statistics, 25(3):573–595, 2013.

[4] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. Basics and Trends in Sensitivity Analysis: Theory and Practice in R. SIAM, 2021.

[5] L. Devroye, L. Györfi, G. Lugosi, and H. Walk. A nearest neighbor estimate of the residual variance. Electronic Journal of Statistics, 12(1):1752–1778, 2018.

[6] L. Devroye, D. Schäfer, L. Györfi, and H. Walk. The estimation problem of minimum mean squared error. Statistics & Decisions, 21(1):15–28, 2003.

[7] F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics. Bernoulli, 28(4):2345–2374, 2022.

[8] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol Pick-Freeze Monte Carlo method. Statistics, 50(4):881–902, 2016.

[9] E. Giné, R. Nickl, et al. A simple adaptive estimator of the integrated square of a density. Bernoulli, 14(1):47–61, 2008.

[10] W. Hoeffding. A class of statistics with asymptotically normal distribution. Ann. Math. Statistics, 19:293–325, 1948.

[11] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol index estimators. ESAIM: Probability and Statistics, 18:342–364, 1 2014.

[12] K. Pearson. On the partial correlation ratio. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 91(632):492–498, 1915.

[13] A. Saltelli, K. Chan, and E. Scott. Sensitivity analysis. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000.

[14] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. Math. Modeling Comput. Experiment, 1(4):407–414 (1995), 1993.