11th International Conference on Sensitivity Analysis of Model Output, April 23–25 2025, Grenoble, France

Recent advances in the understanding and implementation of the HSIC-ANOVA decomposition

Gabriel SARAZIN
Université Paris-Saclay, CEA/DES/ISAS/DM2S/SGLS, 91191 Gif-sur-Yvette, France.

Amandine MARREL
CEA/DES/IRESNE/DER/SESI, 13108 Saint-Paul-Lez-Durance, France.
LMA Université d'Avignon, EA 2151, 84029 Avignon, France.

Sébastien DA VEIGA
Université de Rennes, ENSAI, CNRS, CREST - UMR 9194, 35000 Rennes, France.

Vincent CHABRIDON
EDF R&D, 6 Quai Watier, 78401 Chatou, France.

For anyone wishing to perfom GSA[1] on the output of a black-box model, the ANOVA[2] framework, relying on the estimation of the first-order and total-order Sobol' indices, appears to be the most enticing solution, since it combines both simplicity and explainability. Indeed, each subset of input variables is assigned a specific share of variance equal to the variance induced by the associated sub-function in the Sobol'-Hoeffding decomposition. Due to their nice mathematical properties, the total-order Sobol' allow to perform both the ranking and screening of input variables, making them appear as some of the most attractive sensitivity measures. Unfortunately, when the output variable is computed by a highly expensive computer code, the simulation budget required to achieve an accurate estimation of Sobol' indices is often prohibitive, unless constructing a surrogate model, which is a challenging task in high dimension. In the light of this problem, the sensitivity measures based on the HSIC[3] offer a great alternative, as they are particularly easy to estimate, even when the available data comes from a small Monte Carlo sample. However, while HSIC indices are well adapted to screening, they are not recommended for ranking purposes, as comparing them to one another is not mathematically rigorous.

In this context, the HSIC-ANOVA approach is a cutting-edge kernel method seeking to strike a harmonious balance between Sobol' and HSIC indices [1]. The key idea of this breakthrough is to handle the input variables with ANOVA kernels (instead of more usual kernels such as the Gaussian ones). This specific choice allows to derive a kernel-based ANOVA decomposition in which the output variance is replaced by the HSIC between the input vector set and the output variable. Unlike standard HSIC indices, for which there is no notion of order, the HSIC-ANOVA decomposition enables the definition of kernel-based sensitivity indices at all orders, particularly at the first and total orders, in the same spirit as Sobol' indices.

To obtain such an ANOVA decomposition, the kernel selected for each input variable must be ANOVA, meaning that is it must satisfy an orthogonality condition with respect to the input marginal distribution. Unfortunately, for most parametric families of distributions encountered in practice, it is pretty hard to find an ANOVA kernel which is also characteristic. The only exception is the standard uniform distribution, for which there are many possible candidates in the literature, including the so-called unanchored Sobolev kernels [1]. In almost all other cases, it is advisable to orthogonalize the Gaussian kernel, but this implies an extra step of numerical integration whose complexity will increase linearly with sample size.

When first introduced, the HSIC-ANOVA decomposition was praised for two main reasons:

(a) the fact that all HSIC-ANOVA terms can be accurately estimated from a single sample of input-output observations, regardless of the dimension of the input space ;

[1] **GSA:** Global Sensitivity Analysis
[2] **ANOVA:** ANalysis Of VAriance
[3] **HSIC:** Hilbert-Schmidt Independence Criterion

(b) the fact that the HSIC-ANOVA measure may be used as a cost function for Shapley values, thus leading to HSIC-Shapley effects, a collection of importance measures combining most expected properties in GSA.

However, a grey area persists around the HSIC-ANOVA framework, hindering its wider adoption as a reference methodology in GSA. In fact, there are two main areas for improvement.

(P1) An obvious limitation of HSIC-ANOVA indices is their lack of interpretability, which is partly due to the fact that the HSIC-ANOVA decomposition is not a direct consequence of the Sobol'-Hoeffding decomposition. In particular, it is not clear which kind of extra information is captured by the total-order indices (compared to their first-order counterparts). This lack of transparency is a serious issue, as engineers are unlikely to apply a methodology without having a thorough understanding of it.

(P2) The question of how to use HSIC-ANOVA indices for screening was not investigated in [1].

Our talk aims to provide some answers to these two problems. For the sake of simplicity, the discussion is limited to the case where the input variables are mutually independent and all follow the standard uniform distribution.

In response to (P1), the first part of the talk will reveal the inner workings of the HSIC-ANOVA methodology and will establish a connection between the kernel feature maps and the dependence patterns captured by the two types of HSIC-ANOVA indices. The key to greater interpretability is to express the HSIC as a sum of squared covariances over the entire collection of random features induced by the input and output kernels. In fact, adopting this viewpoint on the HSIC-ANOVA decomposition allows to clarify which random features are captured at each order. Among other benefits, this change of perspective will guide the construction of analytical test functions for which HSIC-ANOVA interactions are controllable, ranging from negligible to dominant contributions.

In response to (P2), the second part of the talk will promote HSIC-ANOVA indices as a promising solution for kernel-based independence testing. The starting point is to realize that the unanchored Sobolev kernels are characteristic [2]. This ensures both the first-order and total-order indices characterize independence. A straightforward strategy to test independence is to apply existing methods for HSIC indices to the numerators of the first-order HSIC-ANOVA indices, because they are simply HSIC indices computed with ANOVA kernels. Another possible strategy is to develop specific test procedures for the total-order HSIC-ANOVA indices. It will be shown that three different test procedures can be employed, each suited to a specific range of sample sizes. Finally, an extensive simulation study will reveal that testing independence with the total-order indices can be more powerful, especially when HSIC-ANOVA interactions come into play.

References:

[1] S. Da Veiga, "Kernel-based ANOVA decomposition and Shapley effects – Application to global sensitivity analysis", 2021. Preprint under review, available at `https://hal.science/hal-03108628v1`

[2] G. Sarazin, A. Marrel, S. Da Veiga and V. Chabridon, "New insights into the feature maps of Sobolev kernels: Application in global sensitivity analysis", 2023. Preprint under review, available at `https://cea.hal.science/cea-04320711`

Acknowledgments: This research work is part of the SAMOURAI[4] project funded by the French National Research Agency (ANR-20-CE46-0013). We are grateful for this financial support.

[Gabriel SARAZIN; Université Paris-Saclay, CEA/DES/ISAS/DM2S/SGLS]
[`gabriel.sarazin@cea.fr` – `https://www.researchgate.net/profile/Gabriel-Sarazin-2`]

[4]Simulation Analytics and Meta-model-based solutions for Optimization, Uncertainty and Reliability AnalysIs.