



On the coalitional decomposition of parameters of interest

Marouane Il Idrissi, Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, Jean-Michel Loubes

► To cite this version:

Marouane Il Idrissi, Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, Jean-Michel Loubes. On the coalitional decomposition of parameters of interest. *Comptes Rendus. Mathématique*, 2023, 361, pp.1653–1662. 10.5802/crmath.521 . hal-03927476

HAL Id: hal-03927476

<https://hal.science/hal-03927476>

Submitted on 6 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the coalitional decomposition of parameters of interest

Marouane Il Idrissi^{a,b,c,e}, Nicolas Bousquet^{a,b,d}, Fabrice Gamboa^c, Bertrand Iooss^{a,b,c}, Jean-Michel Loubes^c

^aEDF Lab Chatou, 6 Quai Watier, 78401 Chatou, France

^bSINCLAIR AI Lab., Saclay, France

^cInstitut de Mathématiques de Toulouse, 31062 Toulouse, France

^dSorbonne Université, LPSM, 4 place Jussieu, Paris, France

^eCorresponding Author - Email: marouane.il-idrissi@edf.fr

Abstract

Understanding the behavior of a black-box model with probabilistic inputs can be based on the decomposition of a parameter of interest (e.g., its variance) into contributions attributed to each coalition of inputs (i.e., subsets of inputs). In this paper, we produce conditions for obtaining unambiguous and interpretable decompositions of very general parameters of interest. This allows to recover known decompositions, holding under weaker assumptions than stated in the literature.

Keywords: interpretability, sensitivity analysis, combinatorics, probability theory, statistics

1. Introduction and preliminaries

The decomposition of a parameter of interest, also known as a quantity of interest (QoI) in the uncertainty quantification framework, with respect to (w.r.t.) coalitions of covariables is crucial in both the field of sensitivity analysis of numerical models and in explainable artificial intelligence [12]. These decompositions allow to distribute shares of QoI to the inputs of an input-output black-box model. Depending on the QoI, they both allow to better understand the behavior of such models, and to perform post-hoc interpretability [1].

For instance, the well-known Hoeffding-Sobol' decomposition is a particular instance of output variance decomposition, which has been used for both settings [7, 8, 5]. It relies on a unique decomposition of an input-output model in L^2 . Nevertheless, it requires independent covariables [11], but allows to quantify the influence (in terms of percentages of output variance) of each inputs of a black-box model, as well as interaction influence due to coalitions of inputs.

In this paper, the concept of “coalitional decomposition of QoI” is developed, generalizing the idea of the Hoeffding-Sobol' variance decomposition to other types of QoIs, leveraging results from the field of combinatorics. In particular, Rota's extension of the Möbius inversion formula to partially ordered sets [14]). Necessary conditions are presented in order to define coalitional decompositions of abstract QoIs. It is shown, among other QoI decompositions proposed in the litterature, that the Hoeffding-Sobol' decomposition still holds without the need for independent inputs, but its interpretation as interaction effects holds only when input independence is assumed. Furthermore, a quite general point of view is adopted, allowing to define decompositions for a large variety of QoIs.

1.1. Notations and tools

1.1.1. Inputs, model and outputs

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be some probability space. Let, for $i = 1, \dots, d$, $d \in \mathbb{N}^*$, $(E_i, \mathcal{B}(E_i))$ be abstract polish measurable space, i.e., E_i is a separable completely metrizable topological space, and $\mathcal{B}(E_i)$ denotes its associated Borel σ -algebra. Let $D = \{1, \dots, d\}$ and denote by $\mathcal{P}(D)$ its power-set (i.e., the

set of all possible subsets of D , including $\emptyset$). For any $A \subseteq D$, denote the *marginal* measurable spaces $(E_A, \mathcal{E}_A)$, where

$$E_A = \prod_{i \in A} E_i, \quad \mathcal{E}_A = \bigotimes_{i \in A} \mathcal{B}(E_i) = \mathcal{B} \left(\prod_{i \in A} E_i \right),$$

Set also $(E, \mathcal{E}) := (E_D, \mathcal{E}_D)$. Let $X = (X_1, \dots, X_d)^\top$ be an E -valued random vector (i.e., a measurable function $X : \Omega \rightarrow E$), referred to as the *inputs*. Let P_X be the distribution of the inputs. Define the marginal distributions, for each $A \subset D$, as:

$$P_{X_A} = \int_{E_{\bar{A}}} dP_X,$$

where $X_A = (X_i)_{i \in A}$ is the coalition of inputs whose indices are in A (i.e., the subset X_A of X). Further, $\bar{A}$ denotes the complementary set of A in D (i.e., $\bar{A} = D \setminus A$). Additionally, for every $A \subset D$, the conditional distributions $P_{X_A | X_{\bar{A}}}$ are assumed to be regular, and if not uniquely defined, they are chosen to be regular (see [4], Chap. 4).

Let $G : E \rightarrow Z$ be an measurable function. Here Z denotes an abstract polish space. $G(X)$ is the Z -valued random variable, resulting from the composition of G with X . In the following, the function G is referred to as a *model*, meanwhile $G(X)$ is referred to as the *output* of the model. Denote $\mathbb{P}(E)$ the set of all probability distributions on $(E, \mathcal{E})$. $\mathcal{M}(E)$ denotes the set of Z -valued models, i.e., every Z -valued, measurable functions.

Remark 1. *In essence, the random inputs X , and the output $G(X)$ are not restricted to be real-valued, but can be defined on more complex measurable spaces (e.g., images, functions, stochastic processes).*

A particular subset of $\mathcal{M}(E)$ is of interest in the present work whenever $Z = \mathbb{R}$: $L^2(P_X, \mathbb{R})$. It is the set of measurable, $\mathbb{R}$ -valued functions which are square-integrable against P_X . Recall that $L^2(P_X; \mathbb{R})$ is a Hilbert space with the inner product:

$$\forall f, g \in L^2(P_X; \mathbb{R}), \quad \langle f, g \rangle_{L^2} = \int_E f(x)g(x)dP_X(x),$$

and associated norm:

$$\forall f \in L^2(P_X; \mathbb{R}), \quad \|f\|_{L^2}^2 = \int_E f^2(x)dP_X(x).$$

Denote, for any $A \subset D$, $L^2(P_{X_A}; \mathbb{R})$ the Hilbert subspaces of $L^2(P_X; \mathbb{R})$, of square integrable, $\mathcal{E}_A$ -measurable functions. In other words, any $f \in L^2(P_{X_A}; \mathbb{R})$ is a square-integrable function $f : E_A \rightarrow \mathbb{R}$: elements of $L^2(P_{X_A}; \mathbb{R})$ only take $|A| := \text{card}(A)$ inputs. Whenever $Z = \mathbb{R}^k$ for a positive integer k , one can also define the set $L^2(P_X; \mathbb{R}^k)$ accordingly (see [9]).

1.1.2. Some elements of combinatorics and abstract algebra

A *partially ordered set* (poset) is defined as a pair $(\mathcal{S}, \leq)$ where $\mathcal{S}$ is a non-empty set, and $\leq$ is a partial order binary relation on elements of $\mathcal{S}$. A poset $(\mathcal{S}, \leq)$ is said to be *locally finite* if, for any $x, z \in \mathcal{S}$, the sets $\{y \in \mathcal{S} : x \leq y \leq z\}$ (also called *segments* of $\mathcal{S}$) are finite.

A *commutative ring with identity*, is a triplet $(\mathbb{A}, +, \times)$ where $\mathbb{A}$ is a non-empty set, and where $+$ and $\times$ are addition and multiplication operators respectively, which are both associative and commutative on $\mathbb{A}$, $\times$ is distributive w.r.t. $+$ on $\mathbb{A}$, $\mathbb{A}$ contains both an additive and multiplicative identity, but only an additive inverse. A commutative ring with identity that admits a multiplicative inverse is generally called a *field*. In the following, abstract commutative rings with identity are denoted $\mathbb{A}$, and are assumed to be endowed with the usual addition and multiplication operator, unless stated otherwise. For instance, $\mathbb{R}$ is a commutative ring with identity (it is in fact, a field).

Denote $I_{\mathbb{A}}(\mathcal{S})$ the incidence algebra of a locally finite poset $(\mathcal{S}, \leq)$ over a commutative ring with identity $\mathbb{A}$, i.e., the set of functions $f : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{A}$ such that $f(x, y) = 0$ if $x \not\leq y$ (see [18], Definition 1.2.1

p.10). $(I_{\mathbb{A}}(\mathcal{S}), +, *)$ forms an $\mathbb{A}$ -algebra with the usual pointwise addition $+$ and the usual convolution $*$, i.e., for any $f, g \in I_{\mathbb{A}}(\mathcal{S})$, and any $x, z \in \mathcal{S}$ such that the segment $\{y \in \mathcal{S} : x \leq y \leq z\}$ is non-empty,

$$(f * g)(x, z) = \sum_{x \leq y \leq z} f(x, y)g(y, z).$$

The zeta function $\zeta \in I_{\mathbb{A}}(\mathcal{S})$ is convolutional identity of the incidence algebra, and is defined as, $\forall x, y \in \mathcal{S}$:

$$\zeta(x, y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases}$$

The Möbius function, denoted $\mu \in I_{\mathbb{A}}(\mathcal{S})$, in the case of locally finite posets $\mathcal{S}$, is defined as the *inverse of the zeta function for the convolution operator* defined on the incidence algebra of $\mathcal{S}$, and can be computed recursively, for any $x, y \in \mathcal{S}$ with $x \leq y$, as [13]

$$\mu(x, y) = \begin{cases} 1 & \text{if } x = y \\ - \sum_{x \leq z < y} \mu(x, z) & \text{otherwise.} \end{cases}$$

Finally, in the scope of this work, it is important to note that, for the finite set D , the pair $(\mathcal{P}(D), \subseteq)$ where $\subseteq$ denotes the inclusion between sets, forms an locally finite poset.

1.2. Quantity of interest

A QoI (or parameter of interest) is the mapping of a model $G \in \mathcal{M}(E)$ and an input distribution $P_X \in \mathbb{P}(E)$ to a commutative ring with identity $\mathbb{A}$. They can be formally defined as follows:

Definition 1 (Quantity of interest). *An $\mathbb{A}$ -valued QoI on a model G with random inputs $X \sim P_X$, is an application:*

$$\begin{aligned} \phi : \mathbb{P}(E) \times \mathcal{M}(E) &\rightarrow \mathbb{A} \\ P \times H &\mapsto \phi_P(H). \end{aligned}$$

onto G and P_X , i.e., $\phi_{P_X}(G)$.

Whenever $Z = \mathbb{A} = \mathbb{R}$, for inputs $X \sim P_X$ and a model $G \in L^2(P_X; \mathbb{R})$, an example of a QoI on G and X can be the variance of the output:

$$\begin{aligned} \phi_{P_X}(G) &= \int_E \left(G(x) - \int_E G(t) dP_X(t) \right)^2 dP_X(x) \\ &= \|G - \mathbb{E}[G(X)]\|_{L^2}^2 \\ &= \mathbb{V}(G(X)) \end{aligned}$$

Other examples of QoIs on G can be its generalized moments w.r.t. X , the probability that $G(X)$ exceeds a fixed threshold, or a quantile of $G(X)$ given a certain level. This definition of a QoI is very general on purpose. In essence, QoIs can also be random variables. However, for the sake of simplicity, in the remainder of this work, it is assumed that for any model G with inputs $X \sim P_X$, $\phi_{P_X}(G)$ is not random.

1.3. Möbius inversion formula and the Inclusion-Exclusion principle

Originally, the “classic” Möbius inversion formula has been first discovered in the field of number theory by [15]. It provides a particular relation between pairs of arithmetic functions (i.e., defined on the natural numbers). This result has since been extended to locally finite posets, and became one of the main foundational result in the field of combinatorics [16]. This extension, as stated in [14] (Section 3.1.2 p.108) writes as follows:

Theorem 1 (Möbius inversion formula on locally finite posets). *Let $\mathcal{S}$ be any non-empty set and $(\mathcal{S}, \leq)$ form a locally finite poset, where $\leq$ is a binary relation. Let φ and ψ be functions from $\mathcal{S}$ to $\mathbb{A}$. Then, the following equivalence hold:*

$$\varphi(x) = \sum_{y: y \leq x} \psi(y), \quad \forall x \in \mathcal{S} \quad \Longleftrightarrow \quad \psi(x) = \sum_{y: y \leq x} \varphi(y) \mu(y, x), \quad \forall x \in \mathcal{S}.$$

where μ is the Möbius function.

The Möbius function, for certain particular posets, admit a closed form. In particular, on the locally finite poset formed by $(\mathcal{P}(D), \subseteq)$, for any $B \subseteq A \in \mathcal{P}(D)$, the Möbius function writes (see [16], Corollary p.345):

$$\mu(B, A) = (-1)^{|A| - |B|}.$$

It comes from the fact that the poset $(\mathcal{P}(D), \subseteq)$ is a Boolean lattice [16]. It leads to the following result (see [14] Section 3.1.1 p.108).

Corollary 1 (Möbius inversion formula on power-sets). *Let φ and ψ be functions from $\mathcal{P}(D)$ to $\mathbb{A}$. Then the following equivalence holds:*

$$\varphi_A = \sum_{B \subseteq A} \psi_B, \quad \forall A \in \mathcal{P}(D) \quad \Longleftrightarrow \quad \psi_A = \sum_{B \subseteq A} (-1)^{|A| - |B|} \varphi_B, \quad \forall A \in \mathcal{P}(D).$$

Corollary 1 can be seen as a generalization of the *Inclusion-Exclusion principle*. It allows to decompose additive functions $f : \mathcal{S} \rightarrow \mathbb{R}$ where $\mathcal{S}$ is an algebra of sets. It is widely used in probability theory. However, in light of this generalization, two main differences arise when compared to the classical principle:

- Both statements in Corollary 1 are equivalent, whereas for the classical Inclusion-Exclusion principle, the left-hand statement only implies the right-left statement;
- The functions to be decomposed are not restricted to be additive, and valued in $\mathbb{R}$ (or even a field) anymore, but they must only be valued in a commutative ring with identity (or even, in some cases, an Abelian group).

The consequences of these differences allow to easily define *coalitional QoI decompositions*, for a broad range of QoIs, and with minimal assumptions on the model G and the distribution P_X of its inputs.

2. Coalitional decompositions of QoIs

A *coalition* of inputs indexed by $A \in \mathcal{P}(D)$ refers to the subset of E_A -valued random inputs X_A . In its essence, a coalitional QoI decomposition amounts to writing a QoI as a sum of terms indexed by a set $A \in \mathcal{P}(D)$, relative to each subset X_A of inputs. They can be formally defined as follows.

Definition 2 (Coalitional decompositions). *Let $G \in \mathcal{M}(E)$ be a model with E -valued random inputs $X \sim P_X \in \mathbb{P}(E)$, and $\phi_{P_X}(G)$ be an $\mathbb{A}$ -valued QoI. One says that a QoI $\phi_{P_X}(G)$ admits a coalitional decomposition if it can be written as:*

$$\phi_{P_X}(G) = \sum_{A \in \mathcal{P}(D)} \psi_A$$

where $\psi : \mathcal{P}(D) \rightarrow \mathbb{A}$. The right hand side is referred to as the *coalitional decomposition* of $\phi_{P_X}(G)$.

2.1. Main result

It is important to note that there exists infinitely many coalitional decompositions for a fixed QoI. However, the following result leverages Corollary 1 in order to characterize a particular class of coalitional decompositions. Sufficient conditions on ψ are given in order to ensure a coalitional QoI decomposition. Notice that it remarkably involves very limited assumptions on the probabilistic structure P_X and the model G .

Lemma 1 (Möbius decomposition). *Let $G \in \mathcal{M}$ a model with E -valued random inputs $X \sim P_X \in \mathbb{P}(E)$. Let $\phi_{P_X}(G)$ be a QoI on G . Let $\varphi : \mathcal{P}(D) \rightarrow \mathbb{A}$ be a set function such that:*

$$\varphi_D = \phi_{P_X}(G).$$

and $\forall A \in \mathcal{P}(D)$, φ_A is well-defined. Then, $\phi_{P_X}(G)$ admits the following coalitional decomposition:

$$\phi_{P_X}(G) = \sum_{A \in \mathcal{P}(D)} \psi_A, \quad (1)$$

where, $\forall A \subseteq D$,

$$\psi_A = \sum_{B \subseteq A} (-1)^{|A|-|B|} \varphi_B. \quad (2)$$

This particular characterization of the coalitional decomposition of $\phi_{P_X}(G)$ is referred to as its Möbius decomposition.

Proof of Lemma 1. Since, by assumption, φ_A is well defined $\forall A \in \mathcal{P}(D)$, let:

$$\psi_A = \sum_{B \subseteq A} (-1)^{|A|-|B|} \varphi_B, \quad \forall A \in \mathcal{P}(D).$$

By Corollary 1, it is equivalent to:

$$\varphi_A = \sum_{B \subseteq A} \psi_B, \quad \forall A \in \mathcal{P}(D),$$

and, in particular:

$$\varphi_D = \phi_{P_X}(G) = \sum_{A \in \mathcal{P}(D)} \psi_A.$$

□

One can notice from Lemma 1 that, defining a coalitional QoI decomposition amounts to choosing a set function φ such that $\varphi_D = \phi_{P_X}(G)$, with very limited assumptions on both G and the inputs' probability structure P_X (i.e., the well-definition of φ_A , $\forall A \in \mathcal{P}(D)$).

2.2. Desirable coalitional decomposition properties

Some Möbius decompositions can be trivial: take, for instance, $\varphi_A = \phi_{P_X}(G)$, $\forall A \in \mathcal{P}(D)$. In those cases, even if the decomposition hold, ψ_A is not *meaningful*, in the sense that it is not related to the subset of inputs X_A . Hence, to ensure the meaningfulness of a Möbius decompositions, some properties can be desired, as detailed in the following.

Definition 3 (Gradual Möbius decomposition). *Let $G \in \mathcal{M}(E)$ be a model with E -valued random inputs $X \sim P_X \in \mathbb{P}(E)$, and let $\phi_{P_X}(G)$ be an $\mathbb{A}$ -valued QoI on G . Assume that this QoI admits a Möbius decomposition (i.e., it can be written as (1) with (2)). If φ can be written, for any $A \in \mathcal{P}(D)$, as:*

$$\varphi_A = \phi_{P_X}(f_A),$$

where $f_A \in \mathcal{M}(E_A)$ is a Z -valued E_A -measurable function, then the decomposition is said to be gradual.

The term gradual refers to the functions f_A , whose input dimension is increasing with the cardinal of $A \in \mathcal{P}(D)$. It ensures that each φ_A is somewhat linked to the coalition of inputs $X_A \sim P_{X_A}$ through the functions f_A , and subsequently, ψ_A as well.

While graduality ensures a link between each ψ_A and the coalitions of inputs X_B for $B \subseteq A$, one can also be interested in their subsequent interpretation. In the particular case where $\mathbb{A} = \mathbb{R}$, and where the QoI is not random, one natural desirable property would be to interpret this decomposition as shares of QoI.

Definition 4 (Fractional Möbius decomposition). *Let $G \in \mathcal{M}(E)$ be a model with E -valued random inputs $X \sim P_X$, and let $\phi_{P_X}(G)$ be a non-random, non-zero $\mathbb{R}$ -valued QoI. Assume that $\phi_{P_X}(G)$ admits a Möbius decomposition (i.e., it can be written as (1) with (2)). If, $\forall A \in \mathcal{P}(D)$:*

$$\text{sign}(\psi_A) = \text{sign}(\phi_{P_X}(G)),$$

Then the Möbius decomposition of $\phi_{P_X}(G)$ is said to be fractional.

If a Möbius decomposition of $\psi_{P_X}(G)$ is fractional, it ensures that the ratios,

$$\frac{\psi_A}{\phi_{P_X}(G)}, \quad \forall A \in \mathcal{P}(D), \quad (3)$$

are in $[0, 1]$, and subsequently that,

$$\sum_{A \in \mathcal{P}(D)} \frac{\psi_A}{\phi_{P_X}(G)} = 1.$$

Essentially, it means that these ratios can be interpreted as shares of QoI attributed to each possible coalition of inputs.

3. Möbius decompositions for global sensitivity analysis

The Möbius decompositions defined in Lemma 1 are especially useful in the context of global sensitivity analysis [6]. In particular, this result allows to:

- Show that some existing QoI decompositions proposed in the literature are Möbius decompositions, and actually hold with weaker assumptions on P_X and G ;
- Define decompositions of QoIs being valued in commutative ring with identity other than $\mathbb{R}$.

3.1. Variance decomposition

Let $X \sim P_X$ be E -valued random inputs of an $\mathbb{R}$ -valued model $G \in L^2(P_X; \mathbb{R}) \subseteq \mathcal{M}(E)$, and let:

$$\phi_{P_X}(G) = \mathbb{V}(G(X)),$$

be the $\mathbb{R}$ -valued QoI, i.e., the variance of the random output $G(X)$.

Proposition 1 (Variance decomposition). *Let, $\forall A \in \mathcal{P}(D)$:*

$$f_A(X_A) = \mathbb{E}[G(X) \mid X_A],$$

and,

$$\varphi_A = \phi_{P_X}(f_A) = \mathbb{V}(\mathbb{E}[G(X) \mid X_A]).$$

Then, $\mathbb{V}(G(X))$ admits the following gradual Möbius decomposition:

$$\mathbb{V}(G(X)) = \sum_{A \in \mathcal{P}(D)} \psi_A,$$

where, $\forall A \in \mathcal{P}(D)$,

$$\psi_A = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mathbb{V}(\mathbb{E}[G(X) | X_B]).$$

Additionally, if the inputs are independent (i.e., $P_X = \prod_{i=1}^d P_{X_i}$), then this decomposition is also fractional.

Proof of Proposition 1. Since $G \in L^2(P_X; \mathbb{R})$, one has that, $\forall A \in \mathcal{P}(D)$:

$$\mathbb{V}(\mathbb{E}[G(X) | X_A]) < \infty.$$

Moreover, notice that $\varphi_D = \mathbb{V}(G(X))$. Applying Lemma 1 ultimately proves the decomposition. Whenever the inputs X are independent, it is well known that $\psi_A \geq 0, \forall A \in \mathcal{P}(D)$ (see [7]), and since $\mathbb{V}(G(X)) > 0$, the decomposition is thus fractional. $\square$

This result is analogue to the Hoeffding-Sobol' functional analysis-of-variance (FANOVA) [11, 17]. Traditionally, this decomposition is the result of a functional decomposition of the model G when it is assumed to be in $L^2(P_X; \mathbb{R})$, into orthogonal elements, requiring the inputs to be independent. However, as shown above, this decomposition holds even when the inputs are endowed with a dependence structure. However, one can notice that input independence allow the decomposition to be fractional, and hence, in-fine, lets the ratios (i.e., as in (3)) to be interpreted as a percentage of the output's variance attributed to each input coalition.

3.2. Covariance decomposition

Now, let $G : E \rightarrow \mathbb{R}^2$ be a model with a bivariate output. Denote $G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$ and assume that $G \in L^2(P_X; \mathbb{R}^2)$. Let

$$\begin{aligned} \phi_{P_X}(G) &= \langle G_1 - \mathbb{E}[G_1(X)], G_2 - \mathbb{E}[G_2(X)] \rangle_{L^2} \\ &= \text{Cov}(G_1(X), G_2(X)), \end{aligned}$$

in other words, the QoI is the covariance between the two random outputs of the model.

Proposition 2 (Covariance decomposition). *Let, $\forall A \in \mathcal{P}(D)$:*

$$f_A(X_A) = \begin{pmatrix} \mathbb{E}[G_1(X) | X_A] \\ \mathbb{E}[G_2(X) | X_A] \end{pmatrix},$$

and,

$$\begin{aligned} \varphi_A &= \phi_{P_X}(f_A) = \langle \mathbb{E}[G_1(X) | X_A] - \mathbb{E}[G_1(X)], \mathbb{E}[G_2(X) | X_A] - \mathbb{E}[G_2(X)] \rangle_{L^2} \\ &= \text{Cov}(\mathbb{E}[G_1(X) | X_A], \mathbb{E}[G_2(X) | X_A]) \end{aligned}$$

Then, $\phi_{P_X}(G)$ admits the following gradual Möbius decomposition:

$$\text{Cov}(G_1(X), G_2(X)) = \sum_{A \in \mathcal{P}(D)} \psi_A,$$

where, $\forall A \in \mathcal{P}(D)$,

$$\psi_A = \sum_{B \subseteq A} (-1)^{|A|-|B|} \text{Cov}(\mathbb{E}[G_1(X) | X_B], \mathbb{E}[G_2(X) | X_B]).$$

Proof of Proposition 2. Notice that since $G \in L^2(P_X; \mathbb{R}^2)$, $\forall A \in \mathcal{P}(D)$, the quantities

$$\text{Cov}(\mathbb{E}[G_1(X) | X_A], \mathbb{E}[G_2(X) | X_A])$$

are well defined, and that $\varphi_D = \text{Cov}(G_1(X), G_2(X))$. Applying Lemma 1 then leads to the gradual decomposition. $\square$

Whenever $G : E \rightarrow \mathbb{R}^k$, for $k \in \mathbb{N}^*$, the two previous results can be generalized using a covariance matrix decomposition (see [9]). Let $\mathcal{D}_k$ be the set of $(k \times k)$ symmetric semi-definite (positive or negative) matrices with non-zero entries on the diagonal, and where elements on the diagonal have the same sign. Note that the triplet $(\mathcal{D}_k, +, \circ)$ where $+$ denotes the usual element-wise matrix addition and $\circ$ denotes the element-wise (Hadamard) multiplication, forms a commutative ring with identity (if all the entries were non-zero, it would be a field since the Hadamard inverse would always be well-defined). Let Σ be the covariance matrix of the output $G(X) = (G_1(X), \dots, G_k(X))^\top$, defined element-wise, for $i, j = 1, \dots, k$:

$$\Sigma_{ij} = \text{Cov}(G_i(X), G_j(X)).$$

Σ is necessarily semi-definite positive (since it is a covariance matrix) and is in $\mathcal{D}_k$ under the assumption that each element of the output is not constant almost surely. It is then a $\mathcal{D}_k$ -valued QoI, and can be decomposed as follows:

Proposition 3 (Covariance matrix decomposition). *Let, $\forall A \in \mathcal{P}(D)$, the matrices $\Sigma^A \in \mathcal{D}_k$ be defined element-wise as:*

$$\Sigma_{i,j}^A = \text{Cov}(\mathbb{E}[G_i(X) | X_A], \mathbb{E}[G_j(X) | X_A]), \quad i, j = 1, \dots, k.$$

Then, Σ admits the following gradual Möbius decomposition:

$$\Sigma = \sum_{A \in \mathcal{P}(D)} \psi_A,$$

where, $\forall A \in \mathcal{P}(D)$,

$$\psi_A = \sum_{B \subseteq A} (-1)^{|A|-|B|} \Sigma^B.$$

Proof of Proposition 3. Notice that since $G \in L^2(P_X; \mathbb{R}^k)$, Σ^A is well-defined $\forall A \in \mathcal{P}(D)$. Moreover, notice that $\Sigma_D = \Sigma$. Applying Lemma 1 then leads to the decomposition. $\square$

One can notice that, in that setting, decomposing Σ amounts to performing the variance decomposition of Proposition 1 on the diagonal elements, and the covariance decomposition of Proposition 2 on the other elements.

3.3. Mean maximum-mean discrepancy decomposition

Aside from moment-based quantities, more complicated QoIs can also be decomposed. Such quantities can be based on kernel embedding of the model G . One can refer to [6] for additional details. For the sake of completeness, some elements are recalled here.

Let $G \in \mathcal{M}(E)$, be a Z -valued model with inputs $X \sim P_X \in \mathbb{P}(E)$. Denote P_Y the distribution the random output $G(X)$. Moreover, for any $A \in \mathcal{P}(D)$, let the conditional distribution of $G(X)$ given X_A be denoted by $P_{Y|X_A}$. Let $k : Z \times Z \rightarrow \mathbb{R}$ be a kernel associated with a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [2]. Let:

$$\mu_G(t) = \int_Z k(z, t) dP_Y(z) = \int_E k(G(z), t) dP_X(z) = \mathbb{E}[k(G(X), t)]$$

denotes the *kernel mean embedding* of $G(X)$. Moreover, denote:

$$\mu_{G|X}(t) = \mathbb{E}[k(G(X), t) | X] = k(G(X), t).$$

The *maximum-mean discrepancy* between P_Y and $P_{Y|X}$ is given by:

$$\begin{aligned} \text{MMD}^2(P_Y, P_{Y|X}) &= \|\mu_G - \mu_{G|X}\|_{\mathcal{H}}^2 \\ &= \mathbb{E}[\mu_G(G(X))] + \mu_{G|X}(G(X)) - 2\mathbb{E}[k(G(X), G(X))] \end{aligned}$$

One is interested in the QoI defined as the mean MMD, i.e.,

$$\begin{aligned} S^{\text{MMD}} &:= \mathbb{E}[\text{MMD}^2(P_Y, P_{Y|X})] \\ &= \mathbb{E}[\mu_G(G(X))] - \mathbb{E}[k(G(X), G(X))] \end{aligned}$$

Proposition 4. *Let $X \sim P_X$ be E -valued random inputs of a model $G : E \rightarrow Z$. Let $k : Z \times Z \rightarrow \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}$. Assume that k is such that, $\forall A \in \mathcal{P}(D)$:*

$$S_A^{\text{MMD}} := \mathbb{E}_{X_A}[\text{MMD}^2(P_Y, P_{Y|X_A})] < \infty.$$

Then, S^{MMD} admits the following Möbius decomposition:

$$S^{\text{MMD}} = \sum_{A \in \mathcal{P}(D)} \psi_A,$$

where, $\forall A \in \mathcal{P}(D)$,

$$\psi_A = \sum_{B \subseteq A} (-1)^{|A|-|B|} S_B^{\text{MMD}}$$

Proof. By assumption, S_A^{MMD} is well-defined $\forall A \in \mathcal{P}(D)$. Moreover, notice that $S_D^{\text{MMD}} = S^{\text{MMD}}$. Applying Lemma 1 then leads to the decomposition. $\square$

This decomposition, analogous to the one presented in [6], not only holds when the inputs are independent, but also when they are endowed with a dependence structure.

4. Discussion

Traditionally, in the field of global sensitivity analysis, QoI decompositions are defined using a “model-centric” approach. It can be summarized as follows: find a suitable coalitional decomposition of the model G in L^2 , such that ϕ_{P_X} becomes an additive map when applied to G . For instance, if the QoI is the variance of the output, orthogonality of the ψ_A is often desired (as defined in Definition 2). The new viewpoint provided by this communication adopts an “input-centric” approach: first define a suitable φ_A (as in (2)), such that it accurately represents the effect of X_A , and then define a suitable decomposition using the reverse implication of the Möbius inversion formula. This approach is analogous to the field of cooperative game theory [3], where φ represents the value function of a cooperative game, and ψ_A are none other than its Harsanyi dividends [10]. The understanding and possible combination of both approaches to find theoretically suitable candidates for φ is the subject of ongoing research.

References

- [1] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020.

- [2] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, Boston, MA, 2004.
- [3] J. M. Bilbao. *Cooperative Games on Combinatorial Structures*, volume 26 of *Theory and Decision Library*. Springer US, Boston, MA, 2000.
- [4] L. Breiman. *Probability*. Society for Industrial and Applied Mathematics, 1992.
- [5] C. Bénése, F. Gamboa, J-M. Loubes, and T. Boissin. Fairness seen as global sensitivity analysis. *Machine Learning*, 2022.
- [6] S. Da Veiga. Kernel-based ANOVA decomposition and Shapley effects - Application to global sensitivity analysis. working paper or preprint, 2021.
- [7] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. *Basics and Trends in Sensitivity Analysis. Theory and Practice in R*. SIAM, 2021.
- [8] T. Fel, R. Cadene, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre. Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. In *Advances in Neural Information Processing Systems*, volume 34, pages 26005–26014, 2021.
- [9] F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. Sensitivity indices for multivariate outputs. *Comptes Rendus Mathématique*, 351(7):307–310, 2013.
- [10] J. C. Harsanyi. A Simplified Bargaining Model for the n-Person Cooperative Game. *International Economic Review*, 4(2):194–220, 1963.
- [11] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.
- [12] B. Iooss, R. Kenett, and P. Secchi. Different Views of Interpretability. In *Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches*, pages 1–20. Springer International Publishing, Cham, 2022.
- [13] J. Kock. From Möbius inversion to renormalisation. *Communications in Number Theory and Physics*, 14(1):171–198, 2020.
- [14] J. P. S. Kung, G-C. Rota, and C. Hung Yan. *Combinatorics: the Rota way*. Cambridge University Press, New York, 2012. OCLC: 1226672593.
- [15] A.F. Möbius. Über eine besondere art von umkehrung der reihen. *Journal für die reine und angewandte Mathematik*, 9:105–123, 1832.
- [16] G-C. Rota. On the foundations of combinatorial theory I. Theory of Möbius Functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964.
- [17] I.M Sobol’. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001.
- [18] Eugene Spiegel and Christopher J. O’Donnell. *Incidence algebras*. Number 206 in Monographs and textbooks in pure and applied mathematics. M. Dekker, New York, 1997.