arXiv:2301.02539v1 [math.ST] 6 Jan 2023

感兴趣参数的联盟分解

Marouane Il Idrissi[a, b, c, e], Nicolas Bousquet[a, b, d], Fabrice Gamboa[c], Bertrand Iooss[a, b, c], Jean-Michel Loubes[c]

[a]EDF Lab Chatou, 6 Quai Watier, 78401 Chatou, France
[b]SINCLAIR AI Lab., Saclay, France[c]Institut de Mathématiques de Toulouse, 31062 Toulouse, France[d]Sorbonne Université, LPSM, 4 place Jussieu, Paris, France[e]Corresponding Author - Email: marouane.il-idrissi@edf.fr

摘要

理解具有概率输入的黑箱模型的行为，可以基于将感兴趣参数（例如其方差）分解为归因于各个输入联盟（即输入子集）的贡献。本文给出了获得含义明确且可解释的通用感兴趣参数分解的条件。这使我们能够在比文献所述更弱的假设下，恢复已知的分解。

关键词： 可解释性、敏感性分析、组合学、概率论、统计学

1. 引言与预备知识

在不确定性量化框架中，感兴趣参数也称为感兴趣量（QoI）。对其按协变量联盟进行分解，对于数值模型敏感性分析领域和可解释人工智能都至关重要 [12]。这些分解能够将感兴趣量份额分配给输入输出黑箱模型的各个输入。根据感兴趣量的不同，它们既有助于更好地理解这类模型的行为，也能用于事后可解释性 [1]。

例如，著名的 Hoeffding-Sobol 分解是输出方差分解的一种特例，已应用于上述两种情形 [7, 8, 5]。它基于对输入输出模型的唯一分解 L^2。然而，它要求协变量相互独立 [11], ，但能够量化黑箱模型各输入的影响（以输出方差的百分比表示），以及由输入联盟产生的交互影响。

本文提出了“感兴趣量的联盟分解”这一概念，将 Hoeffding-Sobol’ 方差分解的思想推广到其他类型的感兴趣量，并借鉴了组合学领域的研究成果。特别是，Rota 将 M¨obius 反演公式推广到了偏序集 [14]）。本文还给出了定义抽象感兴趣量联盟分解所需的必要条件。研究表明，在文献提出的其他感兴趣量分解中，Hoeffding-Sobol’ 分解依然成立，且无需输入相互独立；但只有在假设输入相互独立时，才能将其解释为交互效应。此外，本文采用了相当一般的视角，从而能够为多种感兴趣量定义分解。

1.1. 记号与工具

1.1.1. 输入、模型与输出

设 $(\Omega, \mathcal{F}, \mathbf{P})$ 是某个概率空间。对于 $i = 1, \ldots, d$, $d \in \mathbb{N}^*$，设 $(E_i, \mathcal{B}(E_i))$ 为抽象波兰可测空间，即 E_i 是可分完全可度量化拓扑空间，而 $\mathcal{B}(E_i)$ 表示其相应的博雷尔 σ-代数。设 $D = \{1, \ldots, d\}$，并用 $\mathcal{P}(D)$ 表示其幂集（即

包含 $\emptyset$ 在内的 D 的所有可能子集构成的集合）。对于任意 $A \subseteq D$，将边缘可测空间记为 $(E_A, \mathcal{E}_A)$，其中

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

也设 $(E, \mathcal{E}) := (E_D, \mathcal{E}_D)$。令 $X = (X_1, \ldots, X_d)^\top$ 为一个取值于 E 的随机向量（即一个可测函数 $X : \Omega \to E$），称为输入。令 P_X 表示输入的分布。对每个 $A \subset D$，定义边际分布如下：

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

其中，$X_A = (X_i)_{i \in A}$ 是索引属于 A （即子集 X_A，其属于 X）的输入联盟。此外，A 表示 A 在 D 中的补集（即 $A = D \setminus A$）。另外，对于每个 $A \subset D$，假定条件分布 $P_{X_A | X_A}$ 是正则的；若其定义不唯一，则选取正则版本（参见 [4], 第 4 章）。

令 G: $E \to Z$ 为一个可测函数。这里，Z 表示一个抽象波兰空间。$G(X)$ 是取值于 Z 的随机变量，由 G 与 X 复合而成。在下文中，函数G 称为模型；$G(X)$ 称为该模型的输出。用 $\mathbb{P}(E)$ 表示 $(E, \mathcal{E})$ 上所有概率分布的集合。$\mathcal{M}(E)$ 表示取值于 Z 的模型集合，即所有取值于 Z 的可测函数。

备注 1。本质上，随机输入 X 和输出 $G(X)$ 并不局限于取实值，而是可以定义在更复杂的可测空间上（例如图像、函数、随机过程）。

在本研究中，当 $Z = \mathbb{R}$: $L^2(P_X, \mathbb{R})$ 时，$\mathcal{M}(E)$ 的一个特定子集受到关注。它是关于 P_X 平方可积的、取 $\mathbb{R}$ 值的可测函数集合。回顾一下，$L^2(P_X; \mathbb{R})$ 是一个希尔伯特空间，其内积为：

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

以及相应的范数：

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

对于任意 $A \subset D$，记 $L^2(P_{X_A}; \mathbb{R})$ 为 $L^2(P_X; \mathbb{R})$ 中由平方可积的、$\mathcal{E}_A$-可测函数构成的希尔伯特子空间。换言之，任意 $f \in L^2(P_{X_A}, \mathbb{R})$ 都是一个平方可积函数 f: $E_A \to \mathbb{R}$: $L^2(P_{X_A}, \mathbb{R})$ 的元素仅取 $|A| := \mathrm{card}(A)$ 个输入。当 $Z = \mathbb{R}^k$ 为正整数 k 时，也可以定义集合 $L^2\left(P_X; \mathbb{R}^k\right)$ 相应地（见 [9]）。

1.1.2. 组合学与抽象代数的一些内容

一个偏序集（poset）定义为一对 $(\mathcal{S}, \preceq)$，其中 $\mathcal{S}$ 是一个非空集合，而 $\preceq$ 是定义在 $\mathcal{S}$元素上的偏序二元关系。若对任意$x, z \in \mathcal{S}$，集合 $\{y \in \mathcal{S} : x \preceq y \preceq z\}$（也称为 $\mathcal{S}$的区间）均为有限集，则称偏序集 $(\mathcal{S}, \preceq)$ 是局部有限的。

一个含幺交换环是一个三元组 $(\mathbb{A}, +, \times)$，其中 $\mathbb{A}$ 是一个非空集合，$+$ 和$\times$ 分别是加法和乘法运算，且二者在 $\mathbb{A}$上均满足结合律和交换律；$\times$ 关于 $\mathbb{A}$上的 $+$ 满足分配律；$\mathbb{A}$ 同时包含加法单位元和乘法单位元，但只有加法逆元。存在乘法逆元的含幺交换环通常称为域。下文中，抽象含幺交换环记作 $\mathbb{A}$，并假定其配备通常的加法和乘法运算，除非另有说明。例如，$\mathbb{R}$ 是一个含幺交换环（事实上，它还是一个域）。

记 $I_{\mathbb{A}}(\mathcal{S})$ 为局部有限偏序集 $(\mathcal{S}, \preceq)$ 上以含幺交换环 $\mathbb{A}$ 为系数的关联代数，即函数集 f: $\mathcal{S} \times \mathcal{S} \to \mathbb{A}$，满足：若 $f(x, y) = 0$，则 $x \not\preceq y$（参见 [18], 定义 1.2.1

第 10 页）。$(I_{\mathbb{A}}(\mathcal{S})\ +,\ *)$ 构成一个 $\mathbb{A}$-代数，其运算为通常的逐点加法 $+$ 和通常的卷积 $*$，也就是说，对于任意 $f,\ g \in I_{\mathbb{A}}(\mathcal{S})$，以及任意满足区间 $\{y \in \mathcal{S} : x \le y \le z\}$ 非空的 $x,\ z \in \mathcal{S}$

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

ζ 函数 $\zeta \in I_{\mathbb{A}}(\mathcal{S})$ 是关联代数中的卷积单位元，定义如下：$\forall x,\ y \in \mathcal{S}$:

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

莫比乌斯函数记作 $\mu \in I_{\mathbb{A}}(\mathcal{S})$。对于局部有限偏序集 $\mathcal{S}$，它定义为关联代数上卷积算子的 ζ 函数的逆元，并且可以递归计算：对于任意满足 $x \le y$ 的 $x,\ y \in \mathcal{S}$，有 [13]

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

最后，在本文的研究范围内，值得注意的是，对于有限集 D，有序对 $(\mathcal{P}(D)\ \subseteq)$（其中 $\subseteq$ 表示集合之间的包含关系）构成一个局部有限偏序集。

1.2. 感兴趣量

感兴趣量（或感兴趣参数）是将模型 $G \in \mathcal{M}(E)$ 和输入分布$P_X \in \mathbb{P}(E)$ 映射到含幺交换环 $\mathbb{A}$ 的映射。它们的形式化定义如下：

定义 1（感兴趣量）。模型 G 在具有随机输入 $X \sim P_X$的情况下，取值于 $\mathbb{A}$的感兴趣量是一个映射：

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

映射到 G和 P_X，即 $\phi_{P_X}(G)$。

当 $Z = \mathbb{A} = \mathbb{R}$ 时，对于输入 $X \sim P_X$和模型 $G \in L^2(P_X; \mathbb{R})$，在 G和 X 上的一个感兴趣量示例是输出的方差：

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

在 G 上的其他感兴趣量示例包括其关于 X的广义矩、$G(X)$ 超过固定阈值的概率，或给定某个水平时 $G(X)$ 的分位数。感兴趣量的定义有意设得非常宽泛。从本质上说，感兴趣量也可以是随机变量。不过，为了简化起见，在本文余下部分中，假设对于任意具有输入 $X \sim P_X$的模型 G，$\phi_{P_X}(G)$ 都不是随机的。

1.3. M¨obius 反演公式与容斥原理

最初，“经典的”M¨obius 逆公式由 [15]首先在数论领域发现。它给出了算术函数对（即定义在自然数上的函数）之间的一种特殊关系。此后，这一结果被推广到局部有限偏序集，并成为组合学领域的主要基础结果之一 [16]。该推广如 [14]（第 3.1.2 节，第 108 页）所述，写成如下形式：

定理 1（M¨obius 逆公式在局部有限偏序集上的形式）。设 $\mathcal{S}$ 为任意非空集合，且（$\mathcal{S}$、$\leq$）构成一个局部有限偏序集，其中 $\leq$ 为一个二元关系。设 φ 和 ψ 为从 $\mathcal{S}$ 到 $\mathbb{A}$的函数。则以下等价关系成立：

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

其中 μ 是莫比乌斯函数。

莫比乌斯¨函数在某些特定偏序集上具有闭式。特别地，对于由 $(\mathcal{P}(D) \subseteq)$ 构成的局部有限偏序集，任意 $B \subseteq A \in \mathcal{P}(D)$，莫比乌斯¨函数可写为（见[16],推论第345页）：

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

这是因为偏序集 $(\mathcal{P}(D) \subseteq)$ 是一个布尔格 [16]。由此可得以下结果（见 [14]第3.1.1节，第108页）。

推论1 (莫比乌斯¨幂集上的反演公式)。设 φ 和 ψ 是从 $\mathcal{P}(D)$ 到 $\mathbb{A}$的函数。则以下等价关系成立：

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

推论1 可视为容斥原理的一种推广。它允许将可加函数分解为 $f:\ \mathcal{S} \to \mathbb{R}$ 其中 $\mathcal{S}$ 是一个集合代数。它在概率论中得到广泛应用。然而，鉴于这一推广，与经典原理相比，主要有两点不同：

- 推论 1 中的两个陈述是等价的，而对于经典的容斥原理，左侧陈述只能推出右侧陈述；
- 待分解的函数不再局限于可加函数，也不再要求其取值属于 $\mathbb{R}$ （甚至域），而只需取值属于含幺交换环（某些情况下甚至属于阿贝尔群）。

这些差异所带来的结果 使得我们能够轻松定义联盟式 *QoI* 分解，适用于范围广泛的 QoI，并且 只需对模型 G 及其输入的分布 P_X 作出最少的假设。

2. 感兴趣量的联盟式分解

由 $A \in \mathcal{P}(D)$ 索引的输入联盟是指 E_A 值随机输入 X_A 的子集。从本质上说，联盟式 QoI 分解就是将一个感兴趣量写成若干项之和，每一项由集合 $A \in \mathcal{P}(D)$ 索引，并对应于输入的一个子集 X_A。其形式化定义如下。

定义 2（联盟式分解）。设 $G \in \mathcal{M}(E)$ 是一个具有 E 值随机输入$X \sim P_X \in \mathbb{P}(E)$的模型，且 $\phi_{P_X}(G)$ 是一个 $\mathbb{A}$ 值感兴趣量。若一个感兴趣量 $\phi_{P_X}(G)$ 可以写成以下形式，则称其具有联盟式分解：

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

其中 $\psi:\ \mathcal{P}(D) \to \mathbb{A}$。右侧称为 $\phi_{P_X}(G)$ 的联盟分解。

2.1. 主要结果

需要注意的是，对于固定的感兴趣量，联盟式分解有无穷多个。然而，以下结果利用推论 1 来刻画一类特定的联盟式分解。为确保得到联盟式 QoI 分解，给出了关于 ψ 的充分条件。值得注意的是，这些条件对概率结构 P_X 和模型 G的假设非常有限。

引理 1 (M¨obius 分解)。设 $G \in \mathcal{M}$ 是一个具有 E 值随机输入 $X \sim P_X \in \mathbb{P}(E)$ 的模型。设 $\phi_{P_X}(G)$ 是定义在 G 上的一个感兴趣量。设 $\varphi : \mathcal{P}(D) \to \mathbb{A}$ 是一个集合函数，满足：

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

且 $\forall A \in \mathcal{P}(D)$ φ_A 是良定义的。那么， $\phi_{P_X}(G)$ 具有如下联盟分解 :

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

其中， $\forall A \subseteq D$,

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

$\phi_{P_X}(G)$ 的这种特定联盟分解刻画称为其 M¨obius 分解。

引理 *1*的证明。 由于根据假设， φ_A 定义良好 $\forall A \in \mathcal{P}(D)$ ，令：

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

根据推论 1，这等价于：

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

特别地：

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

□

由引理 1 可知，定义联盟式 QoI 分解相当于选择一个集合函数 φ ，使得 $\varphi_D = \phi_{P_X}(G)$ ，而对 G 及输入的概率结构所作的假设都非常有限 P_X （即 φ_A、 $\forall A \in \mathcal{P}(D)$ 定义良好）。

2.2. 理想的联盟分解性质

某些 M¨obius 分解可能是平凡的：例如，取 $\varphi_A = \phi_{P_X}(G)$ $\forall A \in \mathcal{P}(D)$。在这些情况下，即使分解成立， ψ_A 也不具有实际意义，因为它与输入的子集 X_A无关。因此，为了确保 M¨obius 分解具有实际意义，可以要求其满足某些性质，具体如下。

定义 *3*（渐进 *Möbius* 分解）。设 $G \in \mathcal{M}(E)$ 是一个模型，其随机输入 $X \sim P_X \in \mathbb{P}(E)$ 取值于 E, 且 $\phi_{P_X}(G)$ 是定义在 G 上、取值于 $\mathbb{A}$ 的感兴趣量。假设该感兴趣量存在 *Möbius* 分解（即可以写成式 *(1)*，其中满足式 *(2)*）。若对于任意 $A \in \mathcal{P}(D)$， φ 可以写为：

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

其中 $f_A \in \mathcal{M}(E_A)$ 是一个取值于 Z 的 E_A-可测函数，则称该分解为渐进分解。

渐进一词指的是函数 f_A，其输入维度随 $A \in \mathcal{P}(D)$ 的基数增加。它确保每个 φ_A 都通过函数 f_A 与输入联盟 $X_A \sim P_{X_A}$ 有一定联系，进而也与 ψ_A 相关。

虽然渐进性确保了每个 ψ_A 与输入联盟 X_B 在 $B \subseteq A$ 下之间存在联系，但人们也可能关心对这些联系的进一步解释。在 $\mathbb{A} = \mathbb{R}$ 且感兴趣量不是随机变量的特殊情形下，一种自然且理想的性质是将该分解解释为感兴趣量份额。

定义 *4*（分数 *Möbius* 分解）。设 $G \in \mathcal{M}$（E）是一个模型，其输入 $X \sim P_X$ 是取值于 E 的随机变量；设 ϕ_{P_X}（G）是一个非随机、非零且取值于 $\mathbb{R}$ 的感兴趣量。假设 ϕ_{P_X}（G）存在 *Möbius* 分解（即可以写成式 *(1)* 的形式，并满足式 *(2)*）。若 $\forall A \in \mathcal{P}$（$D$）：

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

则称 ϕ_{P_X}（G） 的 *Möbius* 分解是分数型的。

若 ψ_{P_X}（G）的 Möbius 分解是分数型的，则可确保这些比率,

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

属于 $[0,1],$，随后则是,

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

本质上，这意味着这些比率可以解释为归因于每种可能的输入联盟的感兴趣量份额。

3. M¨obius 分解与全局敏感性分析

M¨obius 分解在引理 1 中定义，在全局敏感性分析的背景下尤为有用 [6]。具体而言，该结果使我们能够:

- 证明文献中提出的一些现有感兴趣量分解属于 M¨obius 分解，而且在对 P_X 和 G采用更弱假设的情况下仍然成立;
- 定义　取值于含幺交换环的感兴趣量的分解（除 $\mathbb{R}$外）　.

3.1. 方差分解

设 $X \sim P_X$ 是取值于 E 的随机输入，属于取值于 $\mathbb{R}$ 的模型 $G \in L^2(P_X; \mathbb{R}) \subseteq \mathcal{M}(E)$，并设:

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

为取值于 $\mathbb{R}$ 的感兴趣量，即随机输出 $G(X)$ 的方差。

命题1（方差分解）。 设，　$\forall A \in \mathcal{P}(D)$:

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

and,

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

那么，$\mathbb{V}(G(X))$ 可进行如下渐进式 *M¨obius* 分解:

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

其中， $\forall A \in \mathcal{P}(D)$，

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

此外，如果输入相互独立（即 $P_X = \prod_{i=1}^{d} P_{X_i}$），那么该分解也是分数型的。

命题 *1* 的证明。 由于 $G \in L^2(P_X; \mathbb{R})$，可得 $\forall A \in \mathcal{P}(D)$：

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

此外，注意到 $\varphi_D = \mathbb{V}(G(X))$。应用引理 1 即可最终证明该分解。当输入 X 相互独立时，众所周知 $\psi_A \geq 0$，$\forall A \in \mathcal{P}(D)$（见 [7]），且由于$\mathbb{V}(G(X)) > 0$，因此该分解是分数型的。 □

这一结果类似于 Hoeffding-Sobol'函数方差分析（FANOVA） [11, 17]。传统上，这一分解是对模型 G 在假设其属于 $L^2(P_X; \mathbb{R})$ 时进行函数分解所得的结果，分解为正交元素，并要求输入相互独立。然而，正如上文所示，即使输入具有依赖结构，该分解仍然成立。不过，可以看出，输入独立性使该分解成为分数型的，因此最终能够将比率（即式（3）中的比率）解释为归因于各输入联盟的输出方差百分比。

3.2. 协方差分解

现在，令 $G: E \to \mathbb{R}^2$ 为具有二元输出的模型。记 $G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$ 并假设$G \in L^2(P_X; \mathbb{R}^2)$。令

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

i换句话说，感兴趣量是模型的两个随机输出之间的协方差 el.

命题 2 （协方差分解）。 令， $\forall A \in \mathcal{P}(D)$：

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

and,

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

那么，$\phi_{P_X}(G)$ 具有以下渐进 *Möbius* 分解：

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

其中， $\forall A \in \mathcal{P}(D)$，

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

命题 2 的证明。注意，由于 $G \in L^2(P_X;\ \mathbb{R}^2)$、$\forall A \in \mathcal{P}(D)$，这些量

$$\mathrm{Cov}\left(\mathbb{E}\left[G_1(X) \mid X_A\right], \mathbb{E}\left[G_2(X) \mid X_A\right]\right)$$

定义良好，并且 $\varphi_D = \mathrm{Cov}(G_1(X)\ G_2(X))$。应用引理 1 即可得到渐进分解。

□

每当 $G\colon E \to \mathbb{R}^k$ 对于 $k \in \mathbb{N}^*$ 成立时，前两个结果都可以借助协方差矩阵分解推广（见 [9]）。令 $\mathcal{D}_k$ 为由 $(k \times k)$ 对称半定（正或负）矩阵构成的集合，这些矩阵的对角线元素非零，且对角线上的元素符号相同。注意，三元组 $(\mathcal{D}_k, +, \circ)$ 中，$+$ 表示通常的逐元素矩阵加法，$\circ$ 表示逐元素（Hadamard）乘法；该三元组构成一个含幺交换环（如果所有元素均非零，它将是一个域，因为 Hadamard 逆总是定义良好）。令 Σ 为输出 $G(X) = (G_1(X) \ldots, G_k(X))^\top$ 的协方差矩阵，其按元素定义，其中 $i, j = 1, \ldots, k$:

$$\Sigma_{ij} = \mathrm{Cov}\left(G_i(X), G_j(X)\right).$$

Σ 必然是半正定的（因为它是一个协方差矩阵），并且在输出的每个元素几乎必然不为常数的假设下，属于 $\mathcal{D}_k$。因此，它是一个取值于 $\mathcal{D}_k$ 的感兴趣量，可按如下方式分解：

命题 3 (协方差矩阵分解)。令 $\forall A \in \mathcal{P}(D)$，矩阵 $\Sigma^A \in \mathcal{D}_k$ 按元素定义如下：

$$\Sigma^A_{i,j} = \mathrm{Cov}\left(\mathbb{E}\left[G_i(X) \mid X_A\right], \mathbb{E}\left[G_j(X) \mid X_A\right]\right), \quad i, j = 1, \ldots, k.$$

于是，Σ 具有如下渐进 *Möbius* 分解：

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

其中，$\forall A \in \mathcal{P}(D)$,

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

命题 3 的证明。注意，由于 $G \in L^2(P_X;\ \mathbb{R}^k)$，$\Sigma^A$ 是良定义的 $\forall A \in \mathcal{P}(D)$。此外，注意到 $\Sigma_D = \Sigma$。应用引理 1 即可得到该分解。

□

可以注意到，在这种情况下，对 Σ 进行分解，相当于对对角元素应用命题 1 的方差分解，并对其他元素应用命题 2的协方差分解。

3.3. 均值最大均值差异分解

除了基于矩的量之外，更复杂的感兴趣量也可以进行分解。这类量可以基于模型 G 的核嵌入。更多详情请参阅 [6] 。为求完整起见，这里回顾一些要点。

设 $G \in \mathcal{M}(E)$ 是一个以 Z 为值、输入为 $X \sim P_X \in \mathbb{P}(E)$ 的模型。记 P_Y $G(X)$ 这一随机输出的分布。此外，对于任意 $A \in \mathcal{P}(D)$，给定 X_A 时 $G(X)$ 的条件分布记为 $P_{Y|X_A}$。设 $k\colon Z \times Z \to \mathbb{R}$ 是与再生核希尔伯特空间（再生核希尔伯特空间） $\mathcal{H}$ [2] 相关联的核。令：

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

表示 $G(X)$ 的 核均值嵌入 。此外，记：

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

P_Y 与 $P_{Y|X}$ 之间的最大均值差异为：

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

我们关注定义为平均最大均值差异的感兴趣量，即，

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

命题 4。 设 $X \sim P_X$ 为模型 G： $E \to Z$的 E值随机输入。设 k： $Z \times Z \to \mathbb{R}$为再生核希尔伯特空间 $\mathcal{H}$ 的再生核。假设 k 满足， $\forall A \in \mathcal{P}(D)$：

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

那么， S^{MMD} 具有如下 *M¨obius* 分解：

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

其中， $\forall A \in \mathcal{P}(D)$,

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

证明。 根据假设， S_A^{MMD}是良定义的 $\forall A \in \mathcal{P}(D)$。此外，注意到 $S_D^{\mathrm{MMD}} = S^{\mathrm{MMD}}$。应用引理1 即可得到该分解。 □

这种分解与 [6], 中提出的分解类似，不仅适用于输入相互独立的情况，也适用于输入具有依赖结构的情况。

4. 讨论

传统上，在全局敏感性分析领域，感兴趣量分解采用一种“以模型为中心”的方法来定义。该方法可概括如下：为模型 G 在 L^2上找到合适的联盟分解，使得 ϕ_{P_X} 作用于 G时成为加性映射。例如，如果感兴趣量是输出的方差，通常希望 ψ_A 具有正交性（如定义 2所述）。本文提出的新视角采用一种“以输入为中心”的方法：首先定义合适的 φ_A （如（2）所示），使其能够准确表示 X_A的影响，然后利用莫比乌斯反演公式的逆向推论定义合适的分解。这种方法类似于合作博弈论领域， [3], 其中 φ 表示合作博弈的特征函数，而 ψ_A 则是其哈桑尼红利 [10]。如何理解并结合这两种方法，以找到在理论上适用于 φ 的候选方案，是当前研究的课题。

参考文献

[1] A. Barredo Arrieta, N. D´ıaz-Rodr´ıguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Gar- cia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila 和 F. Herrera。《可解释人工智能（XAI）：概念、分类体系、机遇与挑战——迈向负责任的人工智能》。*Information Fusion*，58：82–115，2020年6月。

[2] Alain Berlinet 和 Christine Thomas-Agnan。《概率论与统计学中的再生核希尔伯特空间》。Springer US，波士顿，马萨诸塞州，2004年。[3] J. M. Bilbao。《组合结构上的合作博弈》，理论与决策图书馆第26卷。Springer US，波士顿，马萨诸塞州，2000年。[4] L. Breiman。《概率论》。工业与应用数学学会，1992年。[5] C. B´enesse、F. Gamboa、J-M. Loubes 和 T. Boissin。《将公平性视为全局敏感性分析》。机器学习，2022年。[6] S. Da Veiga。《基于核的方差分析分解与Shapley效应——在全局敏感性分析中的应用》。工作论文或预印本，2021年。[7] S. Da Veiga、F. Gamboa、B. Iooss 和 C. Prieur。《敏感性分析基础与趋势：R语言中的理论与实践》。SIAM，2021年。[8] T. Fel、R. Cadene、M. Chalvidal、M. Cord、D. Vigouroux 和 T. Serre。《关注方差！利用基于Sobol的敏感性分析实现高效黑箱解释》。载于《神经信息处理系统进展》第34卷，第26005–26014页，2021年。[9] F. Gamboa、A. Janon、T. Klein 和 A. Lagnoux。《多元输出的敏感性指数》。《数学报告》，351(7)：307–310，2013年。[10] J. C. Harsanyi。《n人合作博弈的简化讨价还价模型》。《国际经济评论》，4(2)：194–220，1963年。[11] W. Hoeffding。《一类具有渐近正态分布的统计量》。《数学统计年鉴》，19(3)：293–325，1948年。[12] B. Iooss、R. Kenett 和 P. Secchi。《可解释性的不同视角》。载于《工业4.0的可解释性：统计与机器学习方法》，第1–20页。Springer International Publishing，查姆，2022年。[13] J. Kock。《从莫比乌斯反演到重整化》。《数论与物理通讯》，14(1)：171–198，2020年。[14] J. P. S. Kung、G-C. Rota 和 C. Hung Yan。《组合学：Rota之道》。剑桥大学出版社，纽约，2012年。OCLC：1226672593。.. [15] A.F. M¨obius。《论一种特殊的级数求逆方法》。《纯粹与应用数学杂志》，9：105–123，1832年。[16] G-C. Rota。《组合理论基础 I：莫比乌斯函数理论》。《概率论与相关领域》，2(4)：340–368，1964年。[17] I.M Sobol'。《非线性数学模型的全局敏感性指数及其蒙特卡罗估计》。《数学与计算机模拟》，55(1)：271–280，2001年。[18] Eugene Spiegel 和 Christopher J. O' Donnell。《关联代数》。纯粹与应用数学专著与教材第206卷。M. Dekker，纽约，1997年。