Avancées en décomposition de modèles boîtes-noires

Usages et perspectives pour l'interprétabilité des modèles de systèmes critiques

[1]**EDF R&D - Lab Chatou** - PRISME Department
[2]**Institut de Mathématiques de Toulouse**
[3]**SINCLAIR AI Lab**

Workshop SINCLAIR
EDF Lab Saclay, France.
28 Mars 2024

Marouane Il Idrissi[123], Nicolas Bousquet[13], Fabrice Gamboa[2], Bertrand Iooss[123], Jean-Michel Loubes[2]

Contexte

Développement de méthodes d'interprétabilité en apprentissage automatique pour la certification des intelligences artificielles reliées aux systèmes critiques.

Objectif de la présentation :
Discuter des **verrous de compréhension débloqués** pendant la thèse.

Contexte

Développement de méthodes d'interprétabilité en apprentissage automatique pour la certification des intelligences artificielles reliées aux systèmes critiques.

Objectif de la présentation :
Discuter des **verrous de compréhension débloqués** pendant la thèse.

Scope :
Quantification de l'influence d'**entrées** sur une **quantité d'intérêt** (QoI) d'un **modèle**.

Contexte

Développement de méthodes d'interprétabilité en apprentissage automatique pour la certification des intelligences artificielles reliées aux systèmes critiques.

Objectif de la présentation :
Discuter des **verrous de compréhension débloqués** pendant la thèse.

Scope :
Quantification de l'influence d'**entrées** sur une **quantité d'intérêt** (QoI) d'un **modèle**.

Notre position :
Compréhension théorique du mécanisme de **décomposition de QoI/de modèle**.

Contexte

Développement de méthodes d'interprétabilité en apprentissage automatique pour la certification des intelligences artificielles reliées aux systèmes critiques.

Objectif de la présentation :
Discuter des **verrous de compréhension débloqués** pendant la thèse.

Scope :
Quantification de l'influence d'**entrées** sur une **quantité d'intérêt** (QoI) d'un **modèle**.

Notre position :
Compréhension théorique du mécanisme de **décomposition de QoI/de modèle**.

Objectif de la thèse:
Proposer des **méthodes d'interprétabilité maîtrisées et justifiées théoriquement**.
Arguments plus convaincants que la validation empirique pour les autorités de régulation/sûreté.

Contexte

Développement de méthodes d'interprétabilité en apprentissage automatique pour la certification des intelligences artificielles reliées aux systèmes critiques.

Objectif de la présentation :
Discuter des **verrous de compréhension débloqués** pendant la thèse.

Scope :
Quantification de l'influence d'**entrées** sur une **quantité d'intérêt** (QoI) d'un **modèle**.

Notre position :
Compréhension théorique du mécanisme de **décomposition de QoI/de modèle**.

Objectif de la thèse:
Proposer des **méthodes d'interprétabilité maîtrisées et justifiées théoriquement**.
Arguments plus convaincants que la validation empirique pour les autorités de régulation/sûreté.

Au programme :
État des lieux de nos avancées, et proposition de **perspectives prometteuses.**

Décomposer une QoI

Pourquoi décomposer une **quantité d'intérêt** ?

QoI : Prévision ponctuelle, espérance, variance du modèle...

Décomposer une QoI

Pourquoi décomposer une **quantité d'intérêt** ?

QoI : Prévision ponctuelle, espérance, variance du modèle...

Permet de quantifier **l'influence** des entrées d'un modèle quelconque pour :

- Vérifier sa **cohérence** par rapport à l'expertise métier.
- Sélectionner ses entrées avec des **arguments tangibles**.
- Auditer **son équité algorithmique**.
- Mieux comprendre ses **rouages internes**.
- **Prioriser les investissements** pour une meilleure qualité des données.
- Aider à sa **conception**.

Décomposer une QoI

Comment décomposer une **quantité d'intérêt** ?

QoI : Prévision ponctuelle, espérance, variance du modèle...

Décomposer une QoI

Comment décomposer une **quantité d'intérêt** ?

QoI : Prévision ponctuelle, espérance, variance du modèle...

Deux écoles :

IA Explicable (XAI)

Théorie des jeux coopératifs

- Fonctionne pour les **entrées dépendantes**.
- Axiomatique **parlante**, interprétation **obscure**.
- Portée théorique **mal comprise**.

Analyse de sensibilité globale (GSA)

Décomposition de modèle (*HDMR*)

- Réservé aux **entrées indépendantes**.
- Interprétation **intuitive**.
- Cadre théorique **établi et maîtrisé**.

Décomposer une QoI

Comment décomposer une **quantité d'intérêt** ?

QoI : Prévision ponctuelle, espérance, variance du modèle...

Deux écoles :

IA Explicable (XAI)

Théorie des jeux coopératifs

- Fonctionne pour les **entrées dépendantes**.
- Axiomatique **parlante**, interprétation **obscure**.
- Portée théorique **mal comprise**.

Analyse de sensibilité globale (GSA)

Décomposition de modèle (*HDMR*)

- Réservé aux **entrées indépendantes**.
- Interprétation **intuitive**.
- Cadre théorique **établi et maîtrisé**.

Nos travaux :

Ces deux approches **peuvent être formalisées à l'aide d'une méthodologie commune.**

Résoudre leurs **inconvénients** respectifs, et **mieux appréhender leur utilisation**.

Deux notions duales

Notations :

☞ On a d entrées $X = (X_1, \ldots, X_d)$.

☞ $D = \{1, \ldots, d\}$ et $\mathcal{P}_D$ est **l'ensemble des sous-ensembles de** D.

☞ Pour chaque $A \in \mathcal{P}_D$, X_A est un **sous-ensemble des entrées**.

Deux notions duales

Notations :

☞ On a d entrées $X = (X_1, \dots, X_d)$.

☞ $D = \{1, \dots, d\}$ et $\mathcal{P}_D$ est **l'ensemble des sous-ensembles de** D.

☞ Pour chaque $A \in \mathcal{P}_D$, X_A est un **sous-ensemble des entrées**.

Une **mesure d'influence** ϕ associe **un nombre** à chaque **sous-ensemble des entrées**.

En tout, elle peut prendre 2^d valeur possible.

La **mesure d'influence** décompose la **QoI** :

$$\sum_{A \in \mathcal{P}_D} \phi(A) = \textbf{QoI}.$$

Deux notions duales

Notations :

- ☞ On a d entrées $X = (X_1, \ldots, X_d)$.
- ☞ $D = \{1, \ldots, d\}$ et $\mathcal{P}_D$ est **l'ensemble des sous-ensembles de** D.
- ☞ Pour chaque $A \in \mathcal{P}_D$, X_A est un **sous-ensemble des entrées**.

Une **mesure d'influence** ϕ associe **un nombre** à chaque **sous-ensemble des entrées**.
En tout, elle peut prendre 2^d valeur possible.

La **mesure d'influence** décompose la **QoI** :

$$\sum_{A \in \mathcal{P}_D} \phi(A) = \textbf{QoI}.$$

Une **mesure de valeur** v associe **un nombre** à chaque **sous-ensemble des entrées**.
En tout, elle peut prendre 2^d valeur possible aussi.

La **mesure de valeur** évaluée sur **toutes les entrées** est égale à la **QoI** :

$$v(D) = \textbf{QoI}$$

Deux approches

Rappel des conditions :

☞ **Mesure d'influence** : $\sum_{A \in \mathcal{P}_D} \phi(A) = $ **QoI**.

☞ **Mesure de valeur** : $v(D) = $ **QoI**.

Deux approches

Rappel des conditions :

☞ **Mesure d'influence** : $\sum_{A \in \mathcal{P}_D} \phi(A) =$ **QoI**.

☞ **Mesure de valeur** : $v(D) =$ **QoI**.

Il a **deux approches** pour définir une **mesure d'influence**.

Deux approches

Rappel des conditions :

☞ **Mesure d'influence** : $\sum_{A \in \mathcal{P}_D} \phi(A) = \textbf{QoI}$.

☞ **Mesure de valeur** : $v(D) = \textbf{QoI}$.

Il a **deux approches** pour définir une **mesure d'influence**.

Approche mécanique (input-centric)

Somme télescopique

Pour $X = (X_1, X_2)$, on **choisit arbitrairement** v et,

$$v(12) = \underbrace{v(12) - v(1) - v(2)}_{\phi(12)} + \underbrace{v(1)}_{\phi 1} + \underbrace{v(2)}_{\phi(2)} = \textbf{QoI}$$

Pour d entrées, $\forall A \in \mathcal{P}_D$:

$$\phi(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B).$$

Approche canonique (model-centric)

Décomposition de modèle

On décompose de manière **unique** le modèle:

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A),$$

où $G_A(X_A)$ est le **représentant** de X_A.
On en déduit une **mesure d'influence**, $\forall A \in \mathcal{P}_D$:

$$\phi(A) = h\left(G_A(X_A)\right)$$

Théoriquement, ces approches sont liées par **l'inversion de Möbius généralisée** (Rota 1964).

Exemples

Décomposition d'une prévision

Pour un modèle G, et une observation x, on veut décomposer $G(x)$.

Choisissons $v(A) = \mathbb{E}\left[G(X) \mid X_A = x_A\right]$ et la mesure d'influence associée :

$$E_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{E}\left[G(X) \mid X_B = x_B\right].$$

Cette décomposition est **canonique** **si et seulement si** les entrées X sont **mutuellement indépendantes**.

Décomposition de la variance

Pour un modèle G, on veut décomposer $\mathbb{V}(G(X))$.

Choisissons $v(A) = \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_A\right]\right)$ et la mesure d'influence associée :

$$S_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_B\right]\right).$$

Cette décomposition est **canonique** **si et seulement si** les entrées X sont **mutuellement indépendantes**.

Exemples

Décomposition d'une prévision

Pour un modèle G, et une observation x, on veut décomposer $G(x)$.

Choisissons $v(A) = \mathbb{E}\left[G(X) \mid X_A = x_A\right]$ et la mesure d'influence associée :

$$E_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{E}\left[G(X) \mid X_B = x_B\right].$$

Cette décomposition est **canonique** **si et seulement si** les entrées X sont **mutuellement indépendantes**.

Décomposition de la variance

Pour un modèle G, on veut décomposer $\mathbb{V}\left(G(X)\right)$.

Choisissons $v(A) = \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_A\right]\right)$ et la mesure d'influence associée :

$$S_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_B\right]\right).$$

Cette décomposition est **canonique** **si et seulement si** les entrées X sont **mutuellement indépendantes**.

☞ **La structure de dépendance affecte l'interprétation.**

Illustration : Modèle linéaire et entrées Gaussiennes

Prenons le modèle :

$$G(X) = X_1 + X_2 X_3, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right)$$

et décomposons **sa variance** avec la **mesure d'influence**

$$S_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_B\right]\right).$$

Illustration : Modèle linéaire et entrées Gaussiennes

Prenons le modèle :

$$G(X) = X_1 + X_2 X_3, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right)$$

et décomposons **sa variance** avec la **mesure d'influence**

$$S_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_B\right]\right).$$

Entrées dépendantes

Approche mécanique (input-centric)

$$S_1 = 0.5 \quad S_2 = 0, \quad S_3 = \rho^2/2,$$
$$S_{12} = \rho^2/2, \quad S_{13} = -\rho^2/2, \quad S_{23} = 0.5,$$
$$S_{123} = -\rho^2/2$$

☞ **Interprétation n'est pas claire**.

Entrées indépendantes

Approche canonique (model-centric)

$$S_1 = 0.5 \quad S_2 = 0, \quad S_3 = 0,$$
$$S_{12} = 0, \quad S_{13} = 0, \quad S_{23} = 0.5,$$
$$S_{123} = 0$$

☞ **Effets d'interaction de Sobol (2001)**.

Illustration : Modèle linéaire et entrées Gaussiennes

Prenons le modèle :

$$G(X) = X_1 + X_2X_3, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right)$$

et décomposons **sa variance** avec la **mesure d'influence**

$$S_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_B\right]\right).$$

Entrées dépendantes

Approche mécanique (input-centric)

$$S_1 = 0.5 \quad S_2 = 0, \quad S_3 = \rho^2/2,$$
$$S_{12} = \rho^2/2, \quad S_{13} = -\rho^2/2, \quad S_{23} = 0.5,$$
$$S_{123} = -\rho^2/2$$

☞ **Interprétation n'est pas claire**.

Entrées indépendantes

Approche canonique (model-centric)

$$S_1 = 0.5 \quad S_2 = 0, \quad S_3 = 0,$$
$$S_{12} = 0, \quad S_{13} = 0, \quad S_{23} = 0.5,$$
$$S_{123} = 0$$

☞ **Effets d'interaction de Sobol (2001)**.

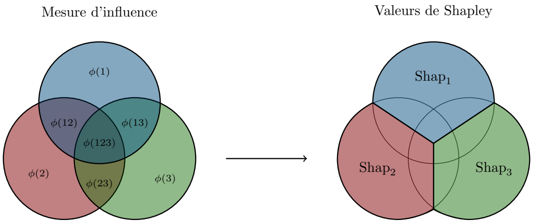
La théorie des jeux coopératifs offre-t-elle une solution ?

Théorie des jeux coopératifs

Les **allocations** ψ décrivent **des agrégations de mesures d'influence**.
On passe de 2^d quantités à d quantités

Les **valeurs de Shapley** sont une répartition **égalitaire** d'une **mesure d'influence** :

$$\text{Shap}_i = \sum_{A \in \mathcal{P}_D : i \in A} \frac{\phi(A)}{|A|}.$$



Quelques remarques :

- **Égalitaire** $\neq$ **Équitable** (fair).
- Il y a **d'autres agrégations** que celle de Shapley.
- Axiomes **régissent uniquement** le processus d'agrégation.

☞ **L'interprétation de l'allocation découle de la mesure d'influence.**

Approche canonique (model-centric)

Quand les entrées sont **mutuellement indépendantes**, on sait décomposer un modèle quelconque

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A),$$

de manière **unique** : C'est la **décomposition d'Hoeffding (1948)**.

☞ Elle **garantie l'interprétation** de la **mesure d'influence** et des **allocations**.

Approche canonique (model-centric)

Quand les entrées sont **mutuellement indépendantes**, on sait décomposer un modèle quelconque

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A),$$

de manière **unique** : C'est la **décomposition d'Hoeffding (1948)**.

☞ Elle **garantie l'interprétation** de la **mesure d'influence** et des **allocations**.

Solution à notre problème :

Généraliser cette décomposition à des **entrées dépendantes**.

Approche canonique (model-centric)

Quand les entrées sont **mutuellement indépendantes**, on sait décomposer un modèle quelconque

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A),$$

de manière **unique** : C'est la **décomposition d'Hoeffding (1948)**.

☞ Elle **garantie l'interprétation** de la **mesure d'influence** et des **allocations**.

Solution à notre problème :

Généraliser cette décomposition à des **entrées dépendantes**.

On est pas les premiers à s'être posé la question.

Rabitz and Aliş (1999) Peccati (2004) Hooker (2007) Kuo et al. (2009) Chastaing, Gamboa, and Prieur (2012) Hart and Gremaud (2018)

Ce que l'on a mis en évidence :

La dépendance introduit des **angles** entre les **espaces de fonctions** des entrées.

☞ Il faut faire des **projections obliques**, et pas orthogonales (espérances conditionnelles).

Quantification d'influence avec entrées dépendantes

Ce que cette généralisation a apporté :

- **Décomposition d'une prévision :**
  - Définir une **mesure d'influence** **unique, canonique et interprétable.**
  - En découle une **mesure de valeur** basée sur des **arguments théorique** pour des **allocations.**

Quantification d'influence avec entrées dépendantes

Ce que cette généralisation a apporté :

- **Décomposition d'une prévision :**
  - Définir une **mesure d'influence** **unique, canonique et interprétable.**
  - En découle une **mesure de valeur** basée sur des **arguments théorique** pour des **allocations.**
- **Décomposition de la variance :**
  - Formaliser ce qu'est un **effet d'interaction**.
  - Formaliser les **effets dûs à la dépendance**.
  - Être en mesure de **distinguer et quantifier les deux**.

Quantification d'influence avec entrées dépendantes

Ce que cette généralisation a apporté :

- **Décomposition d'une prévision :**
  - Définir une **mesure d'influence** **unique, canonique et interprétable.**
  - En découle une **mesure de valeur** basée sur des **arguments théorique** pour des **allocations.**
- **Décomposition de la variance :**
  - Formaliser ce qu'est un **effet d'interaction**.
  - Formaliser les **effets dûs à la dépendance**.
  - Être en mesure de **distinguer et quantifier les deux**.

☞ **Étude éclairée et plus précise de la quantification des incertitudes.**

☞ **Verrou restant : Estimation.**

Quantification d'influence avec entrées dépendantes

Ce que cette généralisation a apporté :

- **Décomposition d'une prévision :**
  - Définir une **mesure d'influence** **unique, canonique et interprétable.**
  - En découle une **mesure de valeur** basée sur des **arguments théorique** pour des **allocations.**
- **Décomposition de la variance :**
  - Formaliser ce qu'est un **effet d'interaction**.
  - Formaliser les **effets dûs à la dépendance**.
  - Être en mesure de **distinguer et quantifier les deux**.

☞ **Étude éclairée et plus précise de la quantification des incertitudes.**

☞ **Verrou restant : Estimation.**

Mais ce n'est pas tout !

La boîte de Pandore est ouverte...

Apprendre le modèle ou apprendre les données ?

Dans un but de **méta-modélisation** : approcher au mieux un modèle $G(X)$ par une fonction $f^{\theta}(X)$.

$G(X)$ et $f_{\theta}(X)$ sont tous les deux décomposables:

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A), \quad f^{\theta}(X) = \sum_{A \in \mathcal{P}_D} f_A^{\theta}(X_A).$$

Question :

Minimiser l'EQM entre $G(X)$ et $f^{\theta}(X)$ $\iff$ Minimiser chaque EQM entre $G_A(X_A)$ et $f_A^{\theta}(X_A)$?

Apprendre le modèle ou apprendre les données ?

Dans un but de **méta-modélisation** : approcher au mieux un modèle $G(X)$ par une fonction $f^{\theta}(X)$.

$G(X)$ et $f_{\theta}(X)$ sont tous les deux décomposables:

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A), \quad f^{\theta}(X) = \sum_{A \in \mathcal{P}_D} f_A^{\theta}(X_A).$$

Question :

Minimiser l'EQM entre $G(X)$ et $f^{\theta}(X)$ $\iff$ Minimiser chaque EQM entre $G_A(X_A)$ et $f_A^{\theta}(X_A)$?

Oui ! Si les entrées sont **mutuellement indépendantes...**
Sinon, on ne sait pas !

Apprendre le modèle ou apprendre les données ?

Dans un but de **méta-modélisation** : approcher au mieux un modèle $G(X)$ par une fonction $f^\theta(X)$.

$G(X)$ et $f_\theta(X)$ sont tous les deux décomposables:

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A), \quad f^\theta(X) = \sum_{A \in \mathcal{P}_D} f_A^\theta(X_A).$$

Question :

Minimiser l'EQM entre $G(X)$ et $f^\theta(X)$ $\Longleftrightarrow$ Minimiser chaque EQM entre $G_A(X_A)$ et $f_A^\theta(X_A)$?

Oui ! Si les entrées sont **mutuellement indépendantes...**
Sinon, on ne sait pas !

☞ **Vers une méthode pour rendre les régressions "interprétables-by-design".**

Étudier la dépendance en profondeur

On **décrit** la dépendance en passant par la notion **d'angles entre espaces fonctionnels**.
Permet de **caractériser la dépendance stochastique entre entrées de différentes nature.**

Question :

Quels sont les liens avec la modélisation "traditionnelle" de la dépendance (copules) ?

Étudier la dépendance en profondeur

On **décrit** la dépendance en passant par la notion **d'angles entre espaces fonctionnels**.
Permet de **caractériser la dépendance stochastique entre entrées de différentes nature.**

Question :
Quels sont les liens avec la modélisation "traditionnelle" de la dépendance (copules) ?

☞ **Vers une meilleure prise en compte et compréhension de la dépendance non-linéaire.**

Comprendre, prendre en compte et formaliser la causalité

La **caractérisation de la dépendance** repose sur la **structure algébrique du treillis Booléen**.
Permet de considérer **toutes** les interactions possibles.

Question :
Encoder et découvrir les liens causaux par le biais d'autres structures algébriques ?

Comprendre, prendre en compte et formaliser la causalité

La **caractérisation de la dépendance** repose sur la **structure algébrique du treillis Booléen**.
Permet de considérer **toutes** les interactions possibles.

Question :
Encoder et découvrir les liens causaux par le biais d'autres structures algébriques ?

☞ **Vers une meilleure prise en compte et un traitement des liens causaux.**
Potentielles synergies avec l'étude des *modèles graphiques*.

Allocations au centre des problématiques métier

Choisir une **mesure de valeur canonique** permet d'interpréter les **allocations**.
Résumer l'information multivariée et non-linéaire est **crucial pour une utilisation pratique**.

Question :
Comment développer des allocations "optimales" et ayant un sens pratique ?

Allocations au centre des problématiques métier

Choisir une **mesure de valeur** **canonique** permet d'interpréter les **allocations**.
Résumer l'information multivariée et non-linéaire est **crucial pour une utilisation pratique**.

Question :
Comment développer des allocations "optimales" et ayant un sens pratique ?

☞ **Vers une standardisation spécifique des méthodes aux domaines d'expertises.**
Potentielles synergies avec le domaine du *transport optimal*.

Conclusion

Ce qu'il faut retenir :

Conclusion

Ce qu'il faut retenir :

- Adopter les IAs pour les systèmes critiques **nécessite une maîtrise théorique des méthodes d'interprétabilité**.

Conclusion

Ce qu'il faut retenir :

- Adopter les IAs pour les systèmes critiques **nécessite une maîtrise théorique des méthodes d'interprétabilité**.
- Il y a **deux approches** pour décomposer une QoI:
  - **Approche mécanique :** "marche" mais n'est **pas toujours interprétable**.
  - **Approche canonique :** permet de **définir des décompositions interprétables**.

Conclusion

Ce qu'il faut retenir :

- Adopter les IAs pour les systèmes critiques **nécessite une maîtrise théorique des méthodes d'interprétabilité**.
- Il y a **deux approches** pour décomposer une QoI:
  - **Approche mécanique :** "marche" mais n'est **pas toujours interprétable**.
  - **Approche canonique :** permet de **définir des décompositions interprétables**.
- Les allocations sont un **résumé (agrégation) de ces décompositions**.

Conclusion

Ce qu'il faut retenir :

- Adopter les IAs pour les systèmes critiques **nécessite une maîtrise théorique des méthodes d'interprétabilité**.
- Il y a **deux approches** pour décomposer une QoI:
  - **Approche mécanique :** "marche" mais n'est **pas toujours interprétable**.
  - **Approche canonique :** permet de **définir des décompositions interprétables**.
- Les allocations sont un **résumé (agrégation) de ces décompositions**.
- La **théorie des jeux coopératifs** est **beaucoup plus riche que les** valeurs de Shapley.

Conclusion

Ce qu'il faut retenir :

- Adopter les IAs pour les systèmes critiques **nécessite une maîtrise théorique des méthodes d'interprétabilité**.
- Il y a **deux approches** pour décomposer une QoI:
  - **Approche mécanique :** "marche" mais n'est **pas toujours interprétable**.
  - **Approche canonique :** permet de **définir des décompositions interprétables**.
- Les allocations sont un **résumé (agrégation) de ces décompositions**.
- La **théorie des jeux coopératifs** est **beaucoup plus riche que les** valeurs de Shapley.
- Généralisation de la décomposition des modèles :
  - Permet de **dissocier les interactions de la dépendance**.
  - Offre une **quantification des incertitudes plus fine**.
  - Offre un **cadre théorique pour étudier les méthodes d'interprétabilité**.
  - Est un outil pour s'attaquer à des **perspectives ambitieuses, riches et variées**.

References i

Axler, S. 2015. *Linear Algebra Done Right* [in en]. Undergraduate Texts in Mathematics. Cham: Springer International Publishing. ISBN: 978-3-319-11079-0 978-3-319-11080-6. https://doi.org/10.1007/978-3-319-11080-6. https://link.springer.com/10.1007/978-3-319-11080-6.

Bryc, W. 1984. "Conditional expectation with respect to dependent sigma-fields." In *Proceedings of VII conference on Probability Theory,* 409–411. https://homepages.uc.edu/~brycwz/preprint/Brasov-1982.pdf.

———. 1996. "Conditional Moment Representations for Dependent Random Variables." Publisher: Institute of Mathematical Statistics and Bernoulli Society, *Electronic Journal of Probability* 1 (none): 1–14. ISSN: 1083-6489, 1083-6489. https://doi.org/10.1214/EJP.v1-7. https://projecteuclid.org/journals/electronic-journal-of-probability/volume-1/issue-none/Conditional-Moment-Representations-for-Dependent-Random-Variables/10.1214/EJP.v1-7.full.

Chastaing, G., F. Gamboa, and C. Prieur. 2012. "Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis." Publisher: Institute of Mathematical Statistics and Bernoulli Society, *Electronic Journal of Statistics* 6, no. none (January): 2420–2448. ISSN: 1935-7524, 1935-7524. https://doi.org/10.1214/12-EJS749. https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-6/issue-none/Generalized-Hoeffding-Sobol-decomposition-for-dependent-variables---application/10.1214/12-EJS749.full.

Da Veiga, S., F. Gamboa, B. Iooss, and C. Prieur. 2021. *Basics and Trends in Sensitivity Analysis. Theory and Practice in R.* SIAM. Computational Science / Engineering.

References ii

Dauxois, J, G. M Nkiet, and Y Romain. 2004. "Canonical analysis relative to a closed subspace." *Linear Algebra and its Applications,* Tenth Special Issue (Part 1) on Linear Algebra and Statistics, 388:119–145. ISSN: 0024-3795. https://doi.org/10.1016/j.laa.2004.02.036. https://www.sciencedirect.com/science/article/pii/S0024379504001107.

Dixmier, J. 1949. "Étude sur les variétés et les opérateurs de Julia, avec quelques applications" [in fre]. *Bulletin de la Société Mathématique de France* 77:11–101. http://eudml.org/doc/86830.

Feshchenko, I. 2020. *When is the sum of closed subspaces of a Hilbert space closed?* https://doi.org/10.48550/arXiv.2012.08688. arXiv: 2012.08688 [math.FA].

Friedrichs, K. 1937. "On Certain Inequalities and Characteristic Value Problems for Analytic Functions and For Functions of Two Variables." Publisher: American Mathematical Society, *Transactions of the American Mathematical Society* 41 (3): 321–364. ISSN: 0002-9947. https://doi.org/10.2307/1989786. https://www.jstor.org/stable/1989786.

Gebelein, H. 1941. "Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung" [in de]. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik* 21 (6): 364–379. ISSN: 00442267, 15214001. https://doi.org/10.1002/zamm.19410210604. https://onlinelibrary.wiley.com/doi/10.1002/zamm.19410210604.

Hart, J., and P. A. Gremaud. 2018. "An approximation theoretic perspective of Sobol' indices with dependent variables" [in English]. Publisher: Begel House Inc. *International Journal for Uncertainty Quantification* 8 (6). ISSN: 2152-5080, 2152-5099. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2018026498. https://www.dl.begellhouse.com/journals/52034eb04b657aea,23dc16a4645b89c9,61d464a51b6bf191.html.

Hoeffding, W. 1948. "A Class of Statistics with Asymptotically Normal Distribution." *The Annals of Mathematical Statistics* 19 (3): 293–325. ISSN: 0003-4851, 2168-8990. https://doi.org/10.1214/aoms/1177730196. https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-19/issue-3/A-Class-of-Statistics-with-Asymptotically-Normal-Distribution/10.1214/aoms/1177730196.full.

Hooker, G. 2007. "Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables" [in en]. *Journal of Computational and Graphical Statistics* 16 (3): 709–732. http://www.jstor.org/stable/27594267.

Joe, H. 1997. *Multivariate Models and Multivariate Dependence Concepts.* New York: Chapman / Hall/CRC. ISBN: 978-0-367-80389-6. https://doi.org/10.1201/9780367803896.

Kallenberg, O. 2021. *Foundations of modern probability.* Probability theory and stochastic modelling. Cham, Switzerland: Springer. ISBN: 978-3-030-61871-1. https://doi.org/10.1007/978-3-030-61871-1.

Koyak, R. A. 1987. "On Measuring Internal Dependence in a Set of Random Variables." Publisher: Institute of Mathematical Statistics, *The Annals of Statistics* 15 (3): 1215–1228. ISSN: 0090-5364, 2168-8966. https://doi.org/10.1214/aos/1176350501. https://projecteuclid.org/journals/annals-of-statistics/volume-15/issue-3/On-Measuring-Internal-Dependence-in-a-Set-of-Random-Variables/10.1214/aos/1176350501.full.

Kuo, F. Y., I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. 2009. "On decompositions of multivariate functions" [in en]. *Mathematics of Computation* 79, no. 270 (November): 953–966. ISSN: 0025-5718. https://doi.org/10.1090/S0025-5718-09-02319-9. http://www.ams.org/journal-getitem?pii=S0025-5718-09-02319-9.

Lebrun, R., and A. Dutfoy. 2009a. “A generalization of the Nataf transformation to distributions with elliptical copula.” *Probabilistic Engineering Mechanics* 24 (2): 172–178. ISSN: 0266-8920. https://doi.org/10.1016/j.probengmech.2008.05.001. https://www.sciencedirect.com/science/article/pii/S0266892008000507.

———. 2009b. “Do Rosenblatt and Nataf isoprobabilistic transformations really differ?” *Probabilistic Engineering Mechanics* 24 (4): 577–584. ISSN: 0266-8920. https://doi.org/10.1016/j.probengmech.2009.04.006. https://www.sciencedirect.com/science/article/pii/S0266892009000307.

Malliavin, P. 1995. *Integration and Probability.* Vol. 157. Graduate Texts in Mathematics. New York, NY: Springer. ISBN: 978-1-4612-8694-3. https://doi.org/10.1007/978-1-4612-4202-4. http://link.springer.com/10.1007/978-1-4612-4202-4.

Mara, Thierry A., Stefano Tarantola, and Paola Annoni. 2015. “Non-parametric methods for global sensitivity analysis of model output with dependent inputs” [in en]. *Environmental Modelling & Software* 72 (October): 173–183. ISSN: 13648152. https://doi.org/10.1016/j.envsoft.2015.07.010. https://linkinghub.elsevier.com/retrieve/pii/S1364815215300153.

Peccati, Giovanni. 2004. “Hoeffding-ANOVA decompositions for symmetric statistics of exchangeable observations.” Publisher: Institute of Mathematical Statistics, *The Annals of Probability* 32 (3): 1796–1829. ISSN: 0091-1798, 2168-894X. https://doi.org/10.1214/009117904000000405. https://projecteuclid.org/journals/annals-of-probability/volume-32/issue-3/Hoeffding-ANOVA-decompositions-for-symmetric-statistics-of-exchangeable-observations/10.1214/009117904000000405.full.

Rabitz, H., and O. Aliş. 1999. “General foundations of high-dimensional model representations” [in en]. *Journal of Mathematical Chemistry* 25 (2): 197–233. ISSN: 1572-8897. https://doi.org/10.1023/A:1019188517934. https://doi.org/10.1023/A:1019188517934.

Rota, G. C. 1964. "On the foundations of combinatorial theory I. Theory of Möbius Functions." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 2 (4): 340–368. ISSN: 1432-2064. https://doi.org/10.1007/BF00531932.

Sidák, Z. 1957. "On Relations Between Strict-Sense and Wide-Sense Conditional Expectations." *Theory of Probability & Its Applications* 2 (2): 267–272. ISSN: 0040-585X. https://doi.org/10.1137/1102020. https://epubs.siam.org/doi/abs/10.1137/1102020.

Sobol, I.M. 2001. "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates." *Mathematics and Computers in Simulation* 55 (1): 271–280. ISSN: 03784754.

Merci pour votre attention !

Des questions ?

More on the generalized Hoeffding decomposition

Random inputs, black-box model

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a **probability space**, and let $(E_1, \mathcal{E}_1), \dots, (E_d, \mathcal{E}_d)$ be **standard Borel measurable spaces**.

The **random inputs** are defined as a **measurable mapping** (i.e., **random element**):

$$X : \Omega \to E,$$

where $E = \bigtimes_{i=1}^{d} E_i$ is the **cartesian product of the d Polish spaces**.

(*This is just a way to say that* $X = (X_1, \dots, X_d)$ *is not necessarily* $\mathbb{R}^d$*-valued*)

> **Remark** . We are mainly going to treat X as **a function**: although **its law is well-defined**, **we don't really need to control it directly**.
>
> (We are going to work with $\mathbb{P}$ instead).

Let $G : E \to \mathbb{R}$ be a **black-box model**, and denote by $G(X)$ the **random output** (it is a random variable).

Generated and $\mathbb{P}$-trivial σ-algebras

Let $D = \{1, \ldots, d\}$, and denote $\mathcal{P}_D$ the **power-set** of D (i.e., the set of subsets of D).

For every $A \subset D$, the **mapping** $X_A = (X_i)_{i \in A}$ is an $E_A := \bigtimes_{i \in A} E_i$-valued **random element**.

For every $A \subset D$, denote by:

- $\sigma_A \subseteq \mathcal{F}$ the **σ-algebra generated by** X_A ;
- $\sigma_X \subseteq \mathcal{F}$ the **σ-algebra generated by** X.

And notice that **if** $B \subseteq A$**, then** $\sigma_B \subseteq \sigma_A$.

> **Lemma** *Doob-Dynkin. If an $\mathbb{R}$-valued random variable Y is σ_X-measurable, then there exists some function $f : E \to \mathbb{R}$ such that $Y = f(X)$ a.s.*

Finally, denote by $\sigma_\emptyset$ the $\mathbb{P}$**-trivial** σ**-algebra**, i.e., the σ-algebra that contains **every event of** $\mathcal{F}$ **of probability** 0.

> **Lemma** *Kallenberg (2021, Lemma 4.9). Every $\sigma_\emptyset$-measurable random variable is a.s. constant.*

Output space

Recall that $(\Omega, \mathcal{F}, \mathbb{P})$ is our sample space, **and let** $\mathcal{G}$ be a **sub-σ-algebra** of $\mathcal{F}$.

Definition *Lebesgue space*. Denote by $\mathbb{L}^2(\mathcal{G})$ the **Lebesgue space** containing every **square-integrable, $\mathbb{R}$-valued random variables**. It is an (infinite-dimensional) Hilbert space with inner product, $\forall Z_1, Z_2 \in \mathbb{L}^2(\mathcal{G})$:

$$\langle Z_1, Z_2 \rangle = \mathbb{E}[Z_1 Z_2] = \int_\Omega Z_1(\omega) Z_2(\omega) d\mathbb{P}(\omega).$$

$\mathbb{L}^2(\sigma_X)$ is the space of **random outputs:** it **only contains random variable that can be expressed as functions of X**.

For every $A \subset D$, $\mathbb{L}^2(\sigma_A) \subset \mathbb{L}^2(\sigma_X)$ **only contains random variables that can be expressed as functions of X_A**.

$\mathbb{L}^2(\sigma_\emptyset)$ **only contains a.s constants**.

Generated Lebesgue subspaces

Theorem *Sidák (1957, Theorem 2)*. Let $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \mathcal{F}$, then

- $\mathbb{L}^2(\mathcal{B}_1) \subseteq \mathbb{L}^2(\mathcal{B}_2) \subseteq \mathbb{L}^2(\mathcal{F})$;
- $\mathbb{L}^2(\mathcal{B}_1) \cap \mathbb{L}^2(\mathcal{B}_2) = \mathbb{L}^2(\mathcal{B}_1 \cap \mathcal{B}_2)$.

Recall that, since for $B \subset A \in \mathcal{P}_D$ we have that $\sigma_B \subseteq \sigma_A$, then:

$\mathbb{L}^2(\sigma_B)$ **is a closed Hilbert subspace of** $\mathbb{L}^2(\sigma_A)$

and all of them are closed subspaces of $\mathbb{L}^2(\sigma_X)$: They are **nested in very a particular way** (more on that later in the talk).

Controlling the Lebesgue spaces w.r.t. the σ-algebras allow to express **spaces of functions of subsets of inputs** (analogously to Chastaing, Gamboa, and Prieur (2012)).

The intuition

Recall the classical result:

Theorem *Malliavin (1995, Chapter 3)*. Let X and Y be two random elements. Then:

$$X \perp\!\!\!\perp Y \iff \forall f(X) \in \mathbb{L}^2(\sigma_X),\ \forall g(Y) \in \mathbb{L}^2(\sigma_Y),\ \mathrm{Corr}\left(f(X), g(Y)\right) = 0,$$

or, in other words, $\mathbb{L}^2_0(\sigma_X) \perp \mathbb{L}^2_0(\sigma_Y)$, where $\mathbb{L}^2_0$ only contains centered random variables.

What does this result entail?

- X and Y are **independent** $\implies$ **The functions of X and Y are uncorrelated.** ✓
- **The functions of X and Y are uncorrelated** $\implies$ X and Y are **independent**. **???**

Intuition:

Is it possible to control the **dependence structure** between the inputs by controlling the **angles between the subspaces** $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A \in \mathcal{P}_D}$?

Dixmier's angle

Definition *Dixmier's angle (Dixmier 1949)*. Let M, N be **closed** subspaces of a Hilbert space H. The cosine of Dixmier's angle between M and N is defined as

$$c_0\left(M, N\right) := \sup\left\{|\langle x, y\rangle| : x \in M, \|x\| \leq 1, \quad y \in N, \|y\| \leq 1\right\}.$$

Dixmier's angle is closely related to the notion of **maximal correlation** in probability theory (Koyak 1987), as a dependence measure between **random elements**.

Definition *Maximal correlation (Gebelein 1941)*. Let Z_1, Z_2 be two **random elements**. The maximal correlation between Z_1 and Z_2 is

$$\rho_0(Z_1, Z_2) := c_0\left(\mathbb{L}_0^2\left(\sigma_{Z_1}\right), \mathbb{L}_0^2\left(\sigma_{Z_2}\right)\right)$$

Remark .

$$Z_1 \perp\!\!\!\perp Z_2 \iff \rho_0\left(Z_1, Z_2\right) = 0.$$

Friedrich's angle

Definition *Friedrich's angle (Friedrichs 1937)*. The cosine of Friedrichs' angle is defined as

$$c\left(M, N\right) := \sup\left\{ |\langle x, y\rangle| : \begin{cases} x \in M \cap \left(M \cap N\right)^{\perp}, \|x\| \leq 1 \\ y \in N \cap \left(M \cap N\right)^{\perp}, \|y\| \leq 1 \end{cases} \right\},$$

where the orthogonal complement is taken w.r.t. to H.

Friedrich's angle is used in probability theory as a measure of **partial dependence** between two random elements (Bryc 1984, 1996; Dauxois, Nkiet, and Romain 2004).

Definition *Maximal partial correlation*. Let Z_1 and Z_2 be two *random elements*. The *maximal partial correlation* is between Z_1 and Z_2 is

$$\rho^*(Z_1, Z_2) := c\left(\mathbb{L}^2\left(\sigma_{Z_1}\right), \mathbb{L}^2\left(\sigma_{Z_2}\right)\right)$$

Remark .

$$\rho^*\left(Z_1, Z_2\right) = 0 \iff \mathbb{E}\left[\mathbb{E}\left[. \mid Z_1\right] \mid Z_2\right] = \mathbb{E}\left[\mathbb{E}\left[. \mid Z_2\right] \mid Z_1\right]$$

Closure and complements

These two angles are related to the **closedness of the sum of the two subspaces**:

- $c\,(M, N) < 1 \iff M + N$ is closed in H ;
- $c_0\,(M, N) < 1 \iff M \cap N = \{0\}$ and $M + N$ is closed in H.

But why should we care?

Because in Hilbert spaces, **a closed subspace is always complemented**, i.e., if M is closed, then there always exists **another subspace** K such that:

$$H = M + K \text{ and } M \cap K = \{0\}.$$

In other words, it makes sense to talk about **"the remainder of the ambient space (*H*) outside of the closed subspace (*M*)".**

One **popular complement** of a closed subspace M is **its orthogonal complement** $M^{\perp}$.

Feshchenko matrix

Let's go back to our set of subspaces $\{\mathbb{L}^2(\sigma_A)\}_{A \in \mathcal{P}_D}$.

How can we "globally" control all the Friedrichs' angles between them?

Intuition: By putting them in a sort of "generalized precision matrix".

Definition *Maximal coalitional precision matrix*. Let Δ be the $(2^d \times 2^d)$, symmetric **set-indexed** matrix, defined element-wise, $\forall A, B \in \mathcal{P}_D$ as

$$\Delta_{AB} = \begin{cases} 1 & \text{if } A = B; \\ -c\left(\mathbb{L}^2(\sigma_A), \mathbb{L}^2(\sigma_B)\right) & \text{otherwise.} \end{cases}$$

These matrices resemble closely the ones used by **Feshchenko (2020)** to study the **closedness of an arbitrary sum of closed subspaces** of a Hilbert space.

$\Longrightarrow$ **We're going to call them "Feshchenko matrices".**

Direct-sum decomposition

An infinite-dimensional **Hilbert space** is still a **linear vector space**.

Definition *Direct-sum decomposition (Axler 2015)*. Let W be a vector space and let $W_1, \ldots, W_n$ be **proper subspaces** of W.

W is said to admit a **direct-sum decomposition** if any $w \in W$ can be written **uniquely** as

$$w = \sum_{i=1}^{n} w_i \text{ where } w_i \in W_i \text{ for } i = 1, \ldots, n.$$

In this case, we write:

$$W = \bigoplus_{i=1}^{n} W_i.$$

<u>Intuition:</u> Can we find a direct-sum decomposition for $\mathbb{L}^2(\sigma_A)$, for every $A \in \mathcal{P}_D$?

If so, we could uniquely decompose any **non-linear function** of X_A, $A \in \mathcal{P}_D$.

But which subspaces should be involved in the direct-sum decomposition ?

Generalized Hoeffding decomposition

Theorem . Under Assumptions **??** and **??**, for every $A \in \mathcal{P}_D$, one has that

$$\mathbb{L}^2(\sigma_A) = \bigoplus_{B \in \mathcal{P}_A} V_B.$$

where $V_\emptyset = \mathbb{L}^2(\sigma_\emptyset)$, and

$$V_B = \left[\sum_{C \in \mathcal{P}_B, C \neq B} V_C \right]^{\perp_B},$$

where $\perp_B$ denotes the orthogonal complement in $\mathbb{L}^2(\sigma_B)$.

Main intuition:

"Inductive generalized centering"

Intuition behind the result: One input

<u>**One input:**</u>

1. Let $i \in D$, and **fix** $\mathbb{L}^2(\sigma_i)$ **as the ambient space**.
2. We have that $V_\emptyset := \mathbb{L}^2(\sigma_\emptyset)$ **is a closed subspace of** $\mathbb{L}^2(\sigma_i)$ (thus it is **complemented**).
3. Denote $V_i = [V_\emptyset]^{\perp_i}$, **the orthogonal complement of** $V_\emptyset$ **in** $\mathbb{L}^2(\sigma_i)$.
4. One has that $\mathbb{L}^2(\sigma_i) = V_\emptyset \oplus V_i$.
5. Since $V_\emptyset$ only contains constants, $V_i = \mathbb{L}_0^2(\sigma_i)$.

In other words, we just showed that any $f(X_i) \in \mathbb{L}^2(\sigma_i)$ can be written as

$$f(X_i) = \underbrace{\mathbb{E}[f(X_i)]}_{\in V_\emptyset} + \underbrace{\mathbb{E}[f(X_i) - \mathbb{E}[f(X_i)]]}_{\in V_i}.$$

And note that we can do this for any $i \in D$.

Intuition behind the result: Two inputs

Two inputs:

1. Let $i, j \in D$, and **fix** $\mathbb{L}^2(\sigma_{ij})$ **as the ambient space**.
2. We have that $\mathbb{L}^2(\sigma_i)$ and $\mathbb{L}^2(\sigma_j)$ are **closed subspaces of** $\mathbb{L}^2(\sigma_{ij})$.
3. **Assumptions ?? and ?? imply that** $\mathbb{L}^2(\sigma_i) + \mathbb{L}^2(\sigma_j)$ **is closed in** $\mathbb{L}^2(\sigma_{ij})$ (thus it is **complemented**).
4. Notice (previous step) that $\mathbb{L}^2(\sigma_i) + \mathbb{L}^2(\sigma_j) = V_\emptyset + V_i + V_j$.
5. Denote $V_{ij} = [V_\emptyset + V_i + V_j]^{\perp_{ij}}$, **the orthogonal complement in** $\mathbb{L}^2(\sigma_{ij})$.
6. We thus have that $\mathbb{L}^2(\sigma_{ij}) = V_\emptyset + V_i + V_j + V_{ij}$.

And note that we can do this for any pair $i, j \in D$**.**

In essence, we **"centered" a bivariate function from its "univariate and constant parts".**

And we can continue the same induction up to d inputs.

Orthocanonical decomposition

As a direct consequence of the previous theorem:

Corollary *Orthocanonical decomposition*. Under Assumptions **??** and **??**, any $G(X) \in \mathbb{L}^2(\sigma_X)$ can be **uniquely decomposed** as

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A),$$

where each $G_A(X_A) \in V_A$.

The term **"orthocanonical"** comes from the choice of the **orthogonal complement** in the "centering process".

The subspaces V_A are comprised of **proper representants**, i.e., either 0 or **functions of exactly** X_A (they do not contain functions of fewer inputs).

Projectors

Recall that for any $G(X) \in \mathbb{L}^2(\sigma_X)$, we have

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A).$$

Oblique projections

Denote the operator

$$Q_A : \mathbb{L}^2(\sigma_X) \to \mathbb{L}^2(\sigma_X), \text{ such that } \quad Q_A(G(X)) = G_A(X_A).$$

Q_A is the (canonical) **oblique projection** onto V_A, parallel to $\bigoplus_{B \in \mathcal{P}_D : B \neq A} V_A$.

Orthogonal projections

Denote the projector

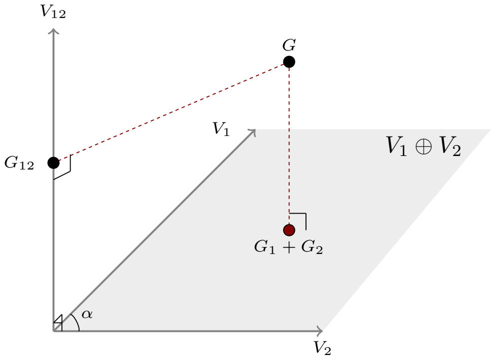
$$P_A : \mathbb{L}^2(\sigma_X) \to \mathbb{L}^2(\sigma_X), \text{ such that } \quad \mathrm{Ran}(P_A) = V_A, \mathrm{Ker}(P_A) = [V_A]^{\perp}.$$

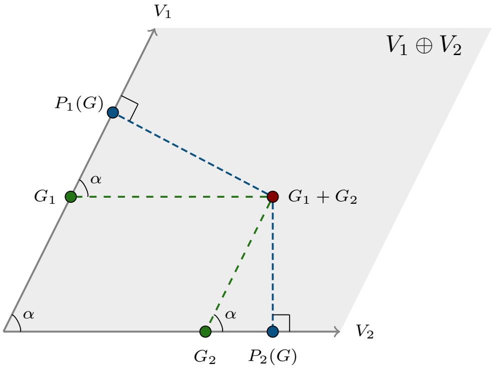
the **orthogonal projection** onto V_A.

Illustration : $\mathbb{L}_0^2(\sigma_{12})$

Hence, for any $G(X) \in \mathbb{L}^2(\sigma_X)$, one has that, $\forall A \in \mathcal{P}_D$

$$G_A(X_A) = Q_A(G(X)).$$





The oblique projection Q_A usually differ from the oblique projections P_A

Oblique and orthogonal projections

In fact,

Proposition . Under Assumptions **??** and **??**,

$$P_A(G(X)) = Q_A(G(X)) \text{ a.s. }, \forall A \in \mathcal{P}_D \quad \iff \quad X \text{ is mutually independent.}$$

This comes from the fact that **the subspaces V_A are all pairwise orthogonal if and only if the inputs are mutually independent.**

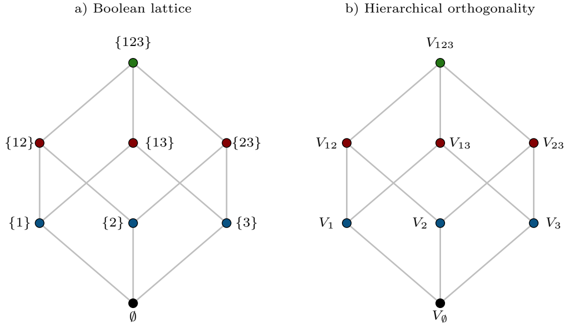
But, under Assumptions **??** and **??**, they **may not be all orthogonal.**

To illustrate this fact, we need some algebraic combinatorics.

Boolean lattice and hierarchical orthogonality

Our decomposition is **over the power-set $\mathcal{P}_D$**, which **which is not trivial**.

When endowed with the **binary relation** $\subseteq$ they form an algebraic structure called **a Boolean lattice**.



The subspaces $\{V_A\}_{A\in\mathcal{P}_D}$ are **hierarchically orthogonal** by design: they follow the same algebraic structure, but this time **w.r.t. to** $\perp$.

More projectors

Recall that:

- Q_A is the **oblique projection** onto V_A.
- P_A is the **orthogonal projection** onto V_A.

But what about projections onto the subspaces $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A\in\mathcal{P}_D}$?

- **(Canonical) oblique projection onto $\mathbb{L}^2(\sigma_A)$:**

$$\begin{aligned}\mathbb{M}_A : \mathbb{L}^2(\sigma_X) &\to \mathbb{L}^2(\sigma_X)\\ G(X) &\mapsto \sum_{B\in\mathcal{P}_A} G_B(X_B)\end{aligned}$$

- **Orthogonal projection onto $\mathbb{L}^2(\sigma_A)$:**

$$\mathbb{E}_A : \mathbb{L}^2(\sigma_X) \to \mathbb{L}^2(\sigma_X), \quad \text{such that } \mathrm{Ran}(\mathbb{E}_A) = \mathbb{L}^2(\sigma_A) \text{ and } \mathrm{Ker}(P_A) = \mathbb{L}^2(\sigma_A)^{\perp},$$

a.k.a **the conditional expectation w.r.t. to X_A** (i.e., $\mathbb{E}[. \mid X_A]$).

Generalized Möbius inversion

Yes, because we're **working on the power-set** $\mathcal{P}_D$!

Corollary *Möbius inversion on power-sets (Rota 1964).* Let $D = \{1, \dots, d\}$. For any two set functions:

$$f : \mathcal{P}_D \to \mathbb{A}, \quad g : \mathcal{P}_D \to \mathbb{A},$$

where $\mathbb{A}$ is an **abelian group**, the following equivalence holds:

$$f(A) = \sum_{B \in \mathcal{P}_A} g(B), \quad \forall A \in \mathcal{P}_D \quad \Longleftrightarrow \quad g(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} f(B), \quad \forall A \in \mathcal{P}_D.$$

In our case, we have, **by definition of the oblique projection onto** $\mathbb{L}^2(\sigma_A)$, that

$$\mathbb{M}_A(G(X)) = \sum_{B \in \mathcal{P}_A} G_B(X_B), \quad \forall A \in \mathcal{P}_D,$$

which is equivalent to

$$G_A(X_A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{M}_B(G(X)), \quad \forall A \in \mathcal{P}_D.$$

Generalized Hoeffding decomposition

If the inputs are mutually independent, from Hoeffding (1948), we have that:

$$G_A(X_A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{E}\left[G(X) \mid X_B\right], \quad \forall A \in \mathcal{P}_D.$$

In our approach, under Assumptions **??** and **??**, we have that:

$$G_A(X_A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{M}_B(G(X)), \quad \forall A \in \mathcal{P}_D.$$

In addition:

Proposition . Under Assumptions **??** and **??**,

$$\mathbb{E}\left[G(X) \mid X_A\right] = \mathbb{M}_A(G(X)) \text{ a.s. }, \forall A \in \mathcal{P}_D \quad \Longleftrightarrow \quad X \text{ is mutually independent.}$$

Our approach actually generalizes Hoeffding's decomposition!

Let's talk about variance decomposition.

We propose **two complementary approaches** for decomposing $\mathbb{V}(G(X))$ based on this generalized decomposition.

Organic variance decomposition: separate **pure interaction effects** to **dependence effects**. The dependence structure of X is **unwanted**, and one wishes to study its effects.

Orthocanonical variance decomposition: the dependence structure of X is **inherent in the uncertainty modeling** of the studied phenomenon. It amounts to quantify **structural** and **correlative** effects.

Organic variance decomposition: Pure interaction

The notion of pure interaction is intrinsically linked with the notion of mutual independence.

Let $\widetilde{X} = (\widetilde{X}_1, \dots, \widetilde{X}_d)^\top$ be the random vector such that

$$\widetilde{X}_i \overset{d}{=} X_i, \qquad \text{and } \widetilde{X} \text{ is mutually independent.}$$

Definition *Pure interaction*. For every $A \in \mathcal{P}_D$, define the **pure interaction of X_A on $G(X)$ as**

$$S_A = \frac{\mathbb{V}\left(P_A(G(\widetilde{X}))\right)}{\mathbb{V}\left(G(\widetilde{X})\right)} \times \mathbb{V}\left(G(X)\right).$$

These indices are the **Sobol' indices** computed on the mutually independent version of X.

This approach **strongly resembles the "independent Sobol' indices"** proposed by Mara, Tarantola, and Annoni (2015).

(see, also, Lebrun and Dutfoy (2009a, 2009b))

Organic variance decomposition: Dependence effects

Recall the following result:

Proposition . Under Assumptions **??** and **??**,

$$P_A(G(X)) = Q_A(G(X)) \text{ a.s. }, \forall A \in \mathcal{P}_D \quad \iff \quad X \text{ is mutually independent.}$$

Which motivates the definition of dependence effects.

Definition *Dependence effects*. For every $A \in \mathcal{P}_D$, define the **dependence effects of X_A on $G(X)$ as**

$$S_A^D = \mathbb{E}\left[\left(Q_A(G(X)) - P_A(G(X))\right)^2 \right].$$

Proposition . Under Assumptions **??** and **??**,

$$S_A^D = 0, \forall A \in \mathcal{P}_D, \quad \iff \quad X \text{ is mutually independent.}$$

What do they sum up to ?...

Probably some interesting global multivariate dependence measure!

Canonical variance decomposition

The structural effects represent the variance of each of the $G_A(X_A)$. It amounts to perform a **covariance decomposition** (Hart and Gremaud 2018; Da Veiga et al. 2021).

Definition *Structural effects*. For every $A \in \mathcal{P}_D$, define the **structural effects of X_A on $G(X)$ as**

$$S_A^U = \mathbb{V}\left(G_A(X_A)\right).$$

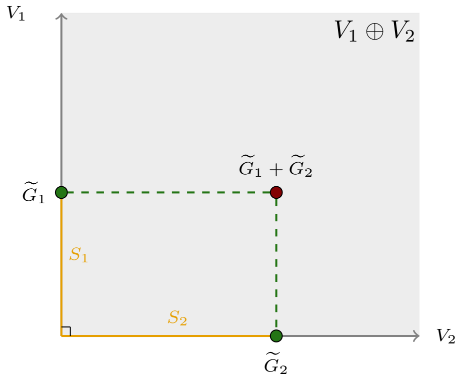
The **correlative effects** represent the part of variance that is due to the correlation between the $G_A(X_A)$.

Definition *Correlative effects*. For every $A \in \mathcal{P}_D$, define the **correlative effects of X_A on $G(X)$ as**

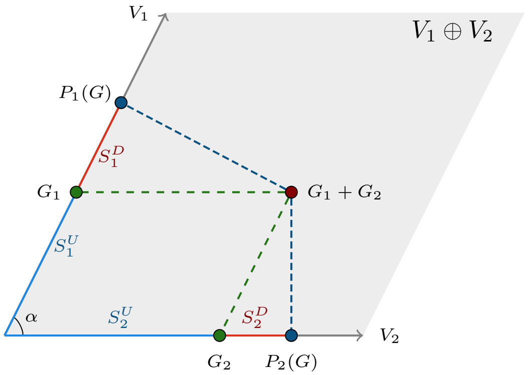
$$S_A^C = \mathrm{Cov}\left(G_A(X_A), \sum_{B \in \mathcal{P}_D : B \neq A} G_B(X_B)\right).$$

Variance decomposition: Intuition

Pure interaction effects



Structural and dependence effects



Example: Two Bernoulli inputs

Let $E = \{0,1\}^2$, and let $X = (X_1, X_2)$, where

$$X_1 \sim \mathcal{B}(q_1), \quad \text{and } X_2 \sim \mathcal{B}(q_2).$$

The joint law of X can be express using **three parameters**:

$$p_{00} = 1 - q_1 - q_2 + \rho, \quad p_{01} = q_2 - \rho, \quad p_{10} = q_1 - \rho, \quad p_{11} = \rho$$

where $p_{ij} = \mathbb{P}(\{X_1 = i\} \cap \{X_2 = j\})$.

Any function $G : \{0,1\}^2 \to \mathbb{R}$ can be expressed as the vector $G = (G_{00}, G_{01}, G_{10}, G_{11})^\top$.

Each value $G_{ij} = G(i,j)$, can be observed with probability p_{ij}.

In this case, we can compute everything analytically.

It requires solving 13 equations with 13 unknowns*.

*https://github.com/milidris/GeneralizedAnova

Feshchenko matrix and the Fréchet bounds

For the **Feshchenko matrix** Δ to be definite positive, one has that:

$$\max\left\{0, q_1q_2 - \sqrt{q_1q_2(1-q_1)(1-q_2)}\right\} < \rho < \min\left\{1, q_1q_2 - \sqrt{q_1q_2(1-q_1)(1-q_2)}\right\}.$$

However, the **classical Fréchet bounds for ρ for bivariate Bernoulli random variables** (Joe 1997, p.210) are equal to

$$\max\left\{0, q_1 + q_2 - 1\right\} \leq \rho \leq \min\left\{q1, q2\right\},$$

and are **more restrictive than the previous ones**.

ρ strictly contained in the Fréchet bounds $\implies$ Assumptions ?? and ?? hold.

Our decomposition hold for virtually any dependence structure between two Bernoullis.