HOEFFDING'S FUNCTIONAL DECOMPOSITION FOR DEPENDENT INPUTS

AND SOME PERSPECTIVES FOR THE INTERPRETATION OF BLACK-BOX MODELS

[1]**Université du Québec à Montréal** (UQÀM)
[2]**Institut Intelligence et Données** (IID) - Universite Laval
[3]**EDF R&D** - PRISME Department
[4]**Institut de Mathématiques de Toulouse**
[5]**SINCLAIR AI Lab**

Conference on New Developments in Probability
Centre de recherches mathématiques - Montréal, QC, Canada.
September 27, 2024

Marouane IL IDRISSI[12(345)], Nicolas BOUSQUET[34], Fabrice GAMBOA[4], Bertrand IOOSS[345], Jean-Michel LOUBES[4]

Introduction

☞ Being able to **understand how black-box models behave under uncertainty** is crucial for **integrating AI models to critical systems**

Industrial processes, energy production, medicine, cars, planes, the stock market...

Introduction

☞ Being able to **understand how black-box models behave under uncertainty** is crucial for **integrating AI models to critical systems**

Industrial processes, energy production, medicine, cars, planes, the stock market...

☞ The decision-making process must be **justifiable** and **justified**

Empirical statements (e.g. SOTA) are notoriously **not enough** to convince **safety/regulation authorities**

Introduction

☞ Being able to **understand how black-box models behave under uncertainty** is crucial for **integrating AI models to critical systems**

Industrial processes, energy production, medicine, cars, planes, the stock market...

☞ The decision-making process must be **justifiable** and **justified**

Empirical statements (e.g. SOTA) are notoriously **not enough** to convince **safety/regulation authorities**

No other choice than relying on theoretical arguments

Introduction

☞ Being able to **understand how black-box models behave under uncertainty** is crucial for **integrating AI models to critical systems**

Industrial processes, energy production, medicine, cars, planes, the stock market...

☞ The decision-making process must be **justifiable** and **justified**

Empirical statements (e.g. SOTA) are notoriously **not enough** to convince **safety/regulation authorities**

No other choice than relying on theoretical arguments

☞ **But**, some of interpretability methods often rely on **unrealistic assumptions**

Usually, mutual independence of the explanatory variables

☞ Being able to **understand how black-box models behave under uncertainty** is crucial for **integrating AI models to critical systems**
Industrial processes, energy production, medicine, cars, planes, the stock market...

☞ The decision-making process must be **justifiable** and **justified**
Empirical statements (e.g. SOTA) are notoriously **not enough** to convince **safety/regulation authorities**

No other choice than relying on theoretical arguments

☞ **But**, some of interpretability methods often rely on **unrealistic assumptions**
Usually, mutual independence of the explanatory variables

This is the case for Hoeffding's classical functional decomposition

Hoeffding's classical decomposition

- ☞ Let $D = \{1, \ldots, d\}$ and let $\mathcal{P}_D$ denote the **power-set** of D
- ☞ Let $X = (X_1, \ldots, X_d)$ be valued in a **cartesian product of Polish spaces**
- ☞ For every $A \in \mathcal{P}_D$, $A \neq \emptyset$, let X_A be a **subset of the inputs**
- ☞ σ_X is the σ-algebra generated by X, σ_A is generated by X_A
- ☞ Let $\mathbb{L}^2(\sigma_X)$ be the **Lebesgue space** of σ_X-measurable random variables
- ☞ Let $G(X) \in \mathbb{L}^2(\sigma_X)$

Hoeffding's classical decomposition

- Let $D = \{1, \dots, d\}$ and let $\mathcal{P}_D$ denote the **power-set** of D
- Let $X = (X_1, \dots, X_d)$ be valued in a **cartesian product of Polish spaces**
- For every $A \in \mathcal{P}_D$, $A \neq \emptyset$, let X_A be a **subset of the inputs**
- σ_X is the σ-algebra generated by X, σ_A is generated by X_A
- Let $\mathbb{L}^2(\sigma_X)$ be the **Lebesgue space** of σ_X-measurable random variables
- Let $G(X) \in \mathbb{L}^2(\sigma_X)$

Hoeffding (1948): For mutually independent X, 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 $G_\emptyset$ is a constant, and the **representants** are all **pairwise orthogonal**, i.e.,

$$\forall A, B \in \mathcal{P}_D, A \neq B, \quad \mathbb{E}\left[G_A(X_A) G_B(X_B)\right] = 0$$

Hoeffding's classical decomposition

- Let $D = \{1, \dots, d\}$ and let $\mathcal{P}_D$ denote the **power-set** of D
- Let $X = (X_1, \dots, X_d)$ be valued in a **cartesian product of Polish spaces**
- For every $A \in \mathcal{P}_D$, $A \neq \emptyset$, let X_A be a **subset of the inputs**
- σ_X is the σ-algebra generated by X, σ_A is generated by X_A
- Let $\mathbb{L}^2(\sigma_X)$ be the **Lebesgue space** of σ_X-measurable random variables
- Let $G(X) \in \mathbb{L}^2(\sigma_X)$

Hoeffding (1948): For mutually independent X, 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 $G_\emptyset$ is a constant, and the **representants** are all **pairwise orthogonal**, i.e.,

$$\forall A, B \in \mathcal{P}_D, A \neq B, \quad \mathbb{E}\left[G_A(X_A) G_B(X_B)\right] = 0$$

Moreover, we can characterize

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

Hoeffding's classical decomposition

☞ The terms $G_A(X_A)$ can be interpreted as **interactions induced by the model** G

This is great for **post-hoc interpretability purposes**!

Hoeffding's classical decomposition

☞ The terms $G_A(X_A)$ can be interpreted as **interactions induced by the model** G
This is great for **post-hoc interpretability purposes**!

For example:

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

In this case, we have that

$$G_1(X_1) = X_1 \quad G_2(X_2) = 0, \quad G_3(X_3) = 0,$$
$$G_{12}(X_{12}) = 0, \quad G_{13}(X_{13}) = 0, \quad G_{23}(X_{23}) = X_2X_3,$$
$$G_{123}(X_{123}) = 0$$

Hoeffding's classical decomposition

☞ The terms $G_A(X_A)$ can be interpreted as **interactions induced by the model** G
This is great for **post-hoc interpretability purposes**!

For example:

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

In this case, we have that

$$G_1(X_1) = X_1 \quad G_2(X_2) = 0, \quad G_3(X_3) = 0,$$
$$G_{12}(X_{12}) = 0, \quad G_{13}(X_{13}) = 0, \quad G_{23}(X_{23}) = X_2X_3,$$
$$G_{123}(X_{123}) = 0$$

We can retrieve the full model by only having access to the **representants**

Hoeffding's classical decomposition

However, the mutual independence assumption is often not respected in practice...

Hoeffding's classical decomposition

However, the mutual independence assumption is often not respected in practice...

In this talk:

We will explore how this result can be generalized to non-mutually independent inputs

Hoeffding's classical decomposition

However, the mutual independence assumption is often not respected in practice...

In this talk:

We will explore how this result can be generalized to non-mutually independent inputs

Tackling such a generalization has already been attempted:

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

Hoeffding's classical decomposition

However, the mutual independence assumption is often not respected in practice...

In this talk:

We will explore how this result can be generalized to non-mutually independent inputs

Tackling such a generalization has already been attempted:

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

But we believe we found an original approach

Defining “non-mutual independence”

☞ Two assumptions on X:

- Non-perfect functional dependence between the variables
  ☞ Conditions on the generated σ-algebras of subsets of X
- Non-degenerate stochastic dependence between the variables
  ☞ Limit the maximal inner product between Lebesgue spaces of subsets of inputs

Interpretation: The variables can be *distinctly recognizable* (at the sample space level)

Defining "non-mutual independence"

☞ Two assumptions on X:

- Non-perfect functional dependence between the variables
  ☞ Conditions on the generated σ-algebras of subsets of X
- Non-degenerate stochastic dependence between the variables
  ☞ Limit the maximal inner product between Lebesgue spaces of subsets of inputs

Interpretation: The variables can be *distinctly recognizable* (at the sample space level)

Non-perfect functional dependence

Assumption 1 *(Non-perfect functional dependence)*.

- $\sigma_\emptyset \subset \sigma_i$, $i = 1, \ldots, d$ (inputs are not constant).
- For $B \subset A$, $\sigma_B \subset \sigma_A$ (inputs add information).
- For every $A, B \in \mathcal{P}_D$, $A \neq B$, $\sigma_A \cap \sigma_B = \sigma_{A \cap B}$.

Non-perfect functional dependence

Assumption 1 *(Non-perfect functional dependence).*

- $\sigma_\emptyset \subset \sigma_i$, $i = 1, \dots, d$ (inputs are not constant).
- For $B \subset A$, $\sigma_B \subset \sigma_A$ (inputs add information).
- For every $A, B \in \mathcal{P}_D$, $A \neq B$, $\sigma_A \cap \sigma_B = \sigma_{A \cap B}$.

☞ $\mathbb{L}^2(\sigma_\emptyset) \subset \mathbb{L}^2(\sigma_A)$, for every $A \in \mathcal{P}_D$

There are non-constant random variables in the Lebesgue spaces

☞ For $B \subset A$, $\mathbb{L}^2(\sigma_B) \subset \mathbb{L}^2(\sigma_A)$

There are functions of X_A that are not functions of X_B

☞ For any $A, B \in \mathcal{P}_D$, $\mathbb{L}^2(\sigma_A) \cap \mathbb{L}^2(\sigma_B) = \mathbb{L}^2(\sigma_{A \cap B})$

The functions of X_A and X_B are functions of $X_{A \cap B}$

Non-perfect functional dependence

Assumption 1 *(Non-perfect functional dependence)*.

- $\sigma_\emptyset \subset \sigma_i$, $i = 1, \ldots, d$ (inputs are not constant).
- For $B \subset A$, $\sigma_B \subset \sigma_A$ (inputs add information).
- For every $A, B \in \mathcal{P}_D$, $A \neq B$, $\sigma_A \cap \sigma_B = \sigma_{A\cap B}$.

☞ $\mathbb{L}^2(\sigma_\emptyset) \subset \mathbb{L}^2(\sigma_A)$, for every $A \in \mathcal{P}_D$

There are non-constant random variables in the Lebesgue spaces

☞ For $B \subset A$, $\mathbb{L}^2(\sigma_B) \subset \mathbb{L}^2(\sigma_A)$

There are functions of X_A that are not functions of X_B

☞ For any $A, B \in \mathcal{P}_D$, $\mathbb{L}^2(\sigma_A) \cap \mathbb{L}^2(\sigma_B) = \mathbb{L}^2(\sigma_{A\cap B})$

The functions of X_A and X_B are functions of $X_{A\cap B}$

Proposition . Under Assumption 1, for any $A, B \in \mathcal{P}_D$ such that $A \cap B \notin \{A, B\}$, **there is no mapping T such that** $X_B = T(X_A)$ a.e.

Non-perfect stochastic dependence

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

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

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

☞ Analogous to the **maximal partial dependence** between random elements (Bryc 1984, 1996; Dauxois, Nkiet, and Romain 2004)

Non-perfect stochastic dependence

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

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

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

☞ Analogous to the **maximal partial dependence** between random elements (Bryc 1984, 1996; Dauxois, Nkiet, and Romain 2004)

Definition *(Feshchenko 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

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

Non-perfect stochastic dependence

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

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

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

☞ Analogous to the **maximal partial dependence** between random elements (Bryc 1984, 1996; Dauxois, Nkiet, and Romain 2004)

Definition *(Feshchenko 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

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

Assumption 2 *(Non-degenerate stochastic dependence).* The Feshchenko matrix Δ of X is definite-positive.

Direct-sum decomposition of $\mathbb{L}^2(\sigma_X)$

Theorem . Under Assumptions 1 and 2, for every $A \in \mathcal{P}_D$,

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

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 $\mathbb{L}^2(\sigma_B)$.

Direct-sum decomposition of $\mathbb{L}^2(\sigma_X)$

Theorem . Under Assumptions 1 and 2, for every $A \in \mathcal{P}_D$,

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

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 $\mathbb{L}^2(\sigma_B)$.

As a consequence, we have the **direct-sum decomposition**:

$$\mathbb{L}^2(\sigma_X) = \bigoplus_{A \in \mathcal{P}_D} V_A,$$

where V_A are **hierarchically orthogonal spaces of pure interactions**

Generalized Hoeffding's decomposition

It implies that any **random output** $G(X) \in \mathbb{L}^2(\sigma_X)$ can be **uniquely** written as:

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

where the **representants** $G_A(X_A) \in V_A$ are **hierarchically orthogonal**

Generalized Hoeffding's decomposition

It implies that any **random output** $G(X) \in \mathbb{L}^2(\sigma_X)$ can be **uniquely** written as:

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

where the **representants** $G_A(X_A) \in V_A$ are **hierarchically orthogonal**

We can characterize the representants by using **oblique projections**

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

Generalized Hoeffding's decomposition

It implies that any **random output** $G(X) \in \mathbb{L}^2(\sigma_X)$ can be **uniquely** written as:

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

where the **representants** $G_A(X_A) \in V_A$ are **hierarchically orthogonal**

We can characterize the representants by using **oblique projections**

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

Recall that Hoeffding (1948) found, **for mutually independent inputs**:

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

Generalized Hoeffding's decomposition

It implies that any **random output** $G(X) \in \mathbb{L}^2(\sigma_X)$ can be **uniquely** written as:

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

where the **representants** $G_A(X_A) \in V_A$ are **hierarchically orthogonal**

We can characterize the representants by using **oblique projections**

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

Recall that Hoeffding (1948) found, **for mutually independent inputs**:

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

Proposition . Under Assumption 1 and 2,

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

Generalized Hoeffding's decomposition

It implies that any **random output** $G(X) \in \mathbb{L}^2(\sigma_X)$ can be **uniquely** written as:

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

where the **representants** $G_A(X_A) \in V_A$ are **hierarchically orthogonal**

We can characterize the representants by using **oblique projections**

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

Recall that Hoeffding (1948) found, **for mutually independent inputs**:

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

Proposition . Under Assumption 1 and 2,

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

We do generalize Hoeffding's decomposition!

More on projectors

In $\mathbb{L}^2(\sigma_X)$, conditional expectations are **orthogonal projections** onto **Lebesgue subspaces**

$\mathbb{E}[. \mid \sigma_A]$ is the orthogonal projection onto $\mathbb{L}^2(\sigma_A)$

More on projectors

In $\mathbb{L}^2(\sigma_X)$, conditional expectations are **orthogonal projections** onto **Lebesgue subspaces**

$\mathbb{E}[. \mid \sigma_A]$ is the orthogonal projection onto $\mathbb{L}^2(\sigma_A)$

The operators $\mathbb{M}_A[.]$ are **oblique projectors** onto the **same Lebesgue subspaces**

$\mathbb{M}_A$ is the canonical projection onto $\mathbb{L}^2(\sigma_A)$ w.r.t. the direct-sum decomposition

More on projectors

In $\mathbb{L}^2(\sigma_X)$, conditional expectations are **orthogonal projections** onto **Lebesgue subspaces**
$\mathbb{E}[. \mid \sigma_A]$ is the orthogonal projection onto $\mathbb{L}^2(\sigma_A)$

The operators $\mathbb{M}_A[.]$ are **oblique projectors** onto the **same Lebesgue subspaces**
$\mathbb{M}_A$ is the canonical projection onto $\mathbb{L}^2(\sigma_A)$ w.r.t. the direct-sum decomposition

☞ **Dependence** introduced **angles between the Lebesgue subspaces** $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A \in \mathcal{P}_D}$
We needed to "skew" the projections accordingly!

More on projectors

In $\mathbb{L}^2(\sigma_X)$, conditional expectations are **orthogonal projections** onto **Lebesgue subspaces**

$\mathbb{E}[. \mid \sigma_A]$ is the orthogonal projection onto $\mathbb{L}^2(\sigma_A)$

The operators $\mathbb{M}_A[.]$ are **oblique projectors** onto the **same Lebesgue subspaces**

$\mathbb{M}_A$ is the canonical projection onto $\mathbb{L}^2(\sigma_A)$ w.r.t. the direct-sum decomposition

☞ **Dependence** introduced **angles between the Lebesgue subspaces** $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A\in\mathcal{P}_D}$

We needed to "skew" the projections accordingly!

Little is known about these oblique projectors in the literature

But they have nice properties (e.g., always commutative)

Main take-aways

☞ Hoeffding's classical result **requires** the explanatory variables to be **mutually independent**
This is limiting in practice...

Main take-aways

☞ Hoeffding's classical result **requires** the explanatory variables to be **mutually independent**
This is limiting in practice...

☞ It can be generalized under two *reasonable* assumptions
Essentially, we need to be able to **distinguish the explanatory variables**

Main take-aways

☞ Hoeffding's classical result **requires** the explanatory variables to be **mutually independent**
This is limiting in practice...

☞ It can be generalized under two *reasonable* assumptions
Essentially, we need to be able to **distinguish the explanatory variables**

☞ It involves "correcting" **conditional expectations** of elements of $\mathbb{L}^2(\sigma_X)$ w.r.t. the dependence structure
Maybe a lot of interesting use to deal with dependence in probability theory!

Main take-aways

☞ Hoeffding's classical result **requires** the explanatory variables to be **mutually independent**
This is limiting in practice...

☞ It can be generalized under two *reasonable* assumptions
Essentially, we need to be able to **distinguish the explanatory variables**

☞ It involves "correcting" **conditional expectations** of elements of $\mathbb{L}^2(\sigma_X)$ w.r.t. the dependence structure
Maybe a lot of interesting use to deal with dependence in probability theory!

☞ The generalized decomposition **paves the way towards a theoretically-grounded interpretation of black-box models**
Many perspectives: algorithmic fairness, causal inference, statistical learning...

Main take-aways

☞ Hoeffding's classical result **requires** the explanatory variables to be **mutually independent**
This is limiting in practice...

☞ It can be generalized under two *reasonable* assumptions
Essentially, we need to be able to **distinguish the explanatory variables**

☞ It involves "correcting" **conditional expectations** of elements of $\mathbb{L}^2(\sigma_X)$ w.r.t. the dependence structure
Maybe a lot of interesting use to deal with dependence in probability theory!

☞ The generalized decomposition **paves the way towards a theoretically-grounded interpretation of black-box models**
Many perspectives: algorithmic fairness, causal inference, statistical learning...

☞ **Main goal: Estimating the oblique projections from data**

Hoeffding decomposition of black-box models with dependent inputs

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

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.

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.

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.

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.

Thank you for your attention!

Any questions?

Non-perfect functional dependence

☞ $\mathbb{L}^2(\sigma_X)$ **contains random variables** **that are functions of** X.

☞ For every $A \subset D$, $\mathbb{L}^2(\sigma_A)$ **contains random variables** **that are functions of** X_A.

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

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

- If $\mathcal{G}_1 \subset \mathcal{G}_2$, then $\mathbb{L}^2(\mathcal{G}_1) \subset \mathbb{L}^2(\mathcal{G}_2) \subseteq \mathbb{L}^2(\mathcal{F})$;
- $\mathbb{L}^2(\mathcal{G}_1) \cap \mathbb{L}^2(\mathcal{G}_2) = \mathbb{L}^2(\mathcal{G}_1 \cap \mathcal{G}_2)$.

Assumption 1 *(Non-perfect functional dependence)*.

- $\sigma_\emptyset \subset \sigma_i$, $i = 1, \dots, d$ (inputs are not constant).
- For $B \subset A$, $\sigma_B \subset \sigma_A$ (inputs add information).
- For every $A, B \in \mathcal{P}_D$, $A \neq B$,

$$\sigma_A \cap \sigma_B = \sigma_{A \cap B}.$$

Output space

Consequences of Assumption 1:

☞ $\mathbb{L}^2(\sigma_\emptyset) \subset \mathbb{L}^2(\sigma_A)$, for every $A \in \mathcal{P}_D$

There are non-constant random variables in the Lebesgue spaces

☞ For $B \subset A$, $\mathbb{L}^2(\sigma_B) \subset \mathbb{L}^2(\sigma_A)$

There are functions of X_A that are not functions of X_B

☞ For any $A, B \in \mathcal{P}_D$,

$$\mathbb{L}^2(\sigma_A) \cap \mathbb{L}^2(\sigma_B) = \mathbb{L}^2(\sigma_{A\cap B})$$

The functions of X_A and X_B are in fact functions of $X_{A\cap B}$

Proposition . Suppose that Assumption 1 hold. Then, for any $A, B \in \mathcal{P}_D$ such that $A \cap B \notin \{A, B\}$, **there is no mapping T such that**

$$X_B = T(X_A) \text{ a.e.}$$

☞ Hence the name **“non-perfect functional dependence”**

Minimal angle, maximal correlation

Dixmier's angle: the **maximal value** the inner product can take **between the elements of two closed subspaces** of a Hilbert space

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

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

☞ Analogous to the **maximal correlation** in probability theory (Koyak 1987), as a dependence measure between **random elements**.

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

$$\rho_0(X, Y) := c_0\left(\mathbb{L}_0^2\left(\sigma_X\right), \mathbb{L}_0^2\left(\sigma_Y\right)\right)$$

Remark . The independence relation from the previous slide can be written as:

$$X \perp\!\!\!\perp Y \iff c_0\left(\mathbb{L}_0^2\left(\sigma_X\right), \mathbb{L}_0^2\left(\sigma_Y\right)\right) = 0.$$

Maximal partial correlation

Friedrichs' angle: Restriction to the **elements orthogonal to the intersection of the subspaces**

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

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

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

☞ Analogous to the **maximal partial dependence** between random elements (Bryc 1984, 1996; Dauxois, Nkiet, and Romain 2004).

Definition *(Maximal partial correlation)*. The **maximal partial correlation** between X and Y is

$$\rho^*(X,Y) := c\left(\mathbb{L}^2(\sigma_X), \mathbb{L}^2(\sigma_Y)\right)$$

Remark . It is closely related **to the commutativity of conditional expectations**.

$$c\left(\mathbb{L}^2(\sigma_X), \mathbb{L}^2(\sigma_Y)\right) = 0 \iff \mathbb{E}\left[\mathbb{E}\left[. \mid X\right] \mid Y\right] = \mathbb{E}\left[\mathbb{E}\left[. \mid Y\right] \mid X\right]$$

Feshchenko matrix

☞ For every $A \in \mathcal{P}_D$, $\mathbb{L}^2(\sigma_A)$ contains **the functions of** X_A
☞ Dixmier's and Friedrichs' angles to **pairwise** control the inner products in these spaces

<u>**Intuition:**</u> A **generalized precision matrix** to control the *global magnitude of all the angles*

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

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

We use **Friedrichs' angles** (partial correlation), hence the **precision** part

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

We're calling them "Feshchenko matrices".

But why are Feshchenko matrices interesting?

Proposition . Suppose that Assumption 1 hold. Then,

$$\Delta = I_{2d} \iff X \text{ is mutually independent.}$$

☞ X is valued in a product of Polish spaces, with **an arbitrary law**

Assumption 2 *(Non-degenerate stochastic dependence)*.

The Feshchenko matrix Δ of the inputs X is definite-positive.

It's a **restriction on the inner product of** $\mathbb{L}^2(\sigma_X) \implies$ **A restriction on the law of** X

☞ Hence the *stochastic dependence* (in opposition to *functional dependence*).

Generalized Hoeffding decomposition

Theorem .

Under Assumptions 1 and 2, 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 $\mathbb{L}^2(\sigma_B)$.

Intuition of the proof:

Inductive functional centering

Intuition of the proof: One input

One input:

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** $\mathbb{L}^2(\sigma_i)$ (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 $\mathbb{L}^2(\sigma_i) = V_\emptyset \oplus V_i$

We just showed that any $f(X_i) \in \mathbb{L}^2(\sigma_i)$ can be written as

$$f(X_i) = \underbrace{\mathbb{E}[f(X_i)]}_{\in V_\emptyset} + \underbrace{\mathbb{E}[f(X_i) - \mathbb{E}[f(X_i)]]}_{\in V_i = \mathbb{L}^2_0(\sigma_i)}$$

And note that $\mathbb{L}^2(\sigma_i) = V_\emptyset \oplus V_i$ **hold for any** $i \in D$ **(induction)**

Intuition of the proof: Two inputs

Two inputs:

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

And note that the decomposition hold for any pair $i, j \in D$

We **"centered" a bivariate function from its "univariate and constant parts"**

We continue the induction up to d inputs.

More projectors

Recall that:

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

But we're more familiar with projections onto $\mathbb{L}^2(\sigma_A)$...
☞ Conditional expectation operators, for example

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

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

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

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

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

Can we characterize Q_A w.r.t. $\mathbb{M}_A$?

Generalized Möbius inversion

Because $(\mathcal{P}_D, \subseteq)$ forms a **Boolean lattice**, yes!

Corollary *(Möbius inversion on power-sets (Rota 1964)).*

For any two set functions:

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

valued in an abelian group $\mathbb{A}$, the following equivalence holds:

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

☞ Analogous to the *inclusion-exclusion principle*

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

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