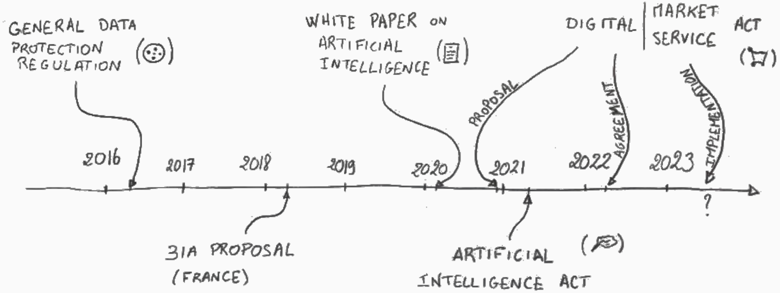
When Global Sensitivity Analysis provides insight into Group Fairness

Clément Bénesse

7 June 2022

IMT - ANITI

Motivation: what society wants for AI



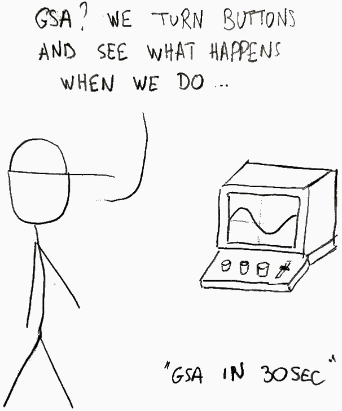
Quote from the *Artificial Intelligence Act* (21/04/2021)

"The measures referred to in paragraph 3 shall enable the individuals to whom human oversight is assigned to do the following, as appropriate to the circumstances:

1. fully understand the capacities and limitations of the high-risk AI system [...] ;
2. remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (‘automation bias’)[...] ;
3. **be able to correctly interpret the high-risk AI system’s output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;**
4. be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;
5. be able to intervene on the operation of the high-risk AI system or interrupt the system through a “stop” button or a similar procedure."

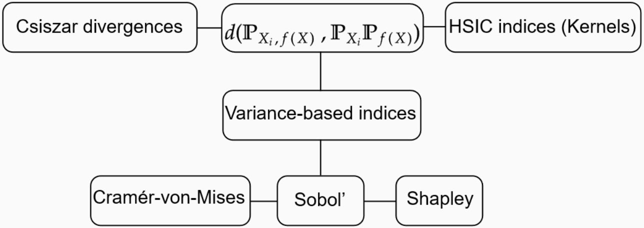
A song of GSA & Fairness

What is GSA?



- GSA = **Global Sensitivity Analysis**
- Quantification of the influence of a variable in a set of input variables $\mathbf{X} := (X_1, \dots, X_p)$ on the outcome of a black-box algorithm f.
- In fact, we want to quantify $d(\mathbb{P}_{(X_i, f(\mathbf{X}))}, \mathbb{P}_{X_i}\mathbb{P}_{f(\mathbf{X})})$, with d a distance for distributions.

Some GSA indices



What are Sobol' indices?

Sobol' indices keywords: Hoeffding decomposition, functional ANOVA.

Assume $\mathbb{P}_{\mathbf{X}} = \prod_{i=1}^{p} \mathbb{P}_{X_i}$ and let $f \in \mathbb{L}^2(\mathbb{P}_{\mathbf{X}}), \mathbb{E}[f] = 0$ (f centered),

$$f(\mathbf{X}) = \sum_{A \in \mathcal{P}(d)} f_A(\mathbf{X}_A),$$

where $\mathbf{X}_A := \{X_i, i \in A\}$ and the $f_A(\mathbf{X}_A) := \sum (-1)^{|A|-|B|} \mathbb{E}[f(\mathbf{X}) | \mathbf{X}_B]$ are orthogonal.

What are Sobol' indices?

Sobol' indices keywords: Hoeffding decomposition, functional ANOVA.

Assume $\mathbb{P}_{\mathbf{X}} = \prod_{i=1}^{p} \mathbb{P}_{X_i}$ and let $f \in \mathbb{L}^2(\mathbb{P}_{\mathbf{X}}), \mathbb{E}[f] = 0$ (f centered),

$$f(\mathbf{X}) = \sum_{A \in \mathcal{P}(d)} f_A(\mathbf{X}_A),$$

where $\mathbf{X}_A := \{X_i, i \in A\}$ and the $f_A(\mathbf{X}_A) := \sum (-1)^{|A|-|B|} \mathbb{E}[f(\mathbf{X})|\mathbf{X}_B]$ are orthogonal.

Then we have:

$$\operatorname{Var} f(\mathbf{X}) = \sum_{A \in \mathcal{P}(d)} \operatorname{Var} f_A(\mathbf{X}_A).$$

What are Sobol' indices?

Sobol' indices keywords: Hoeffding decomposition, functional ANOVA.

Assume $\mathbb{P}_{\mathbf{X}} = \prod_{i=1}^{p} \mathbb{P}_{X_i}$ and let $f \in \mathbb{L}^2(\mathbb{P}_{\mathbf{X}}), \mathbb{E}[f] = 0$ (f centered),

$$f(\mathbf{X}) = \sum_{A \in \mathcal{P}(d)} f_A(\mathbf{X}_A),$$

where $\mathbf{X}_A := \{X_i, i \in A\}$ and the $f_A(\mathbf{X}_A) := \sum (-1)^{|A|-|B|} \mathbb{E}[f(\mathbf{X}) | \mathbf{X}_B]$ are orthogonal.

Then we have:

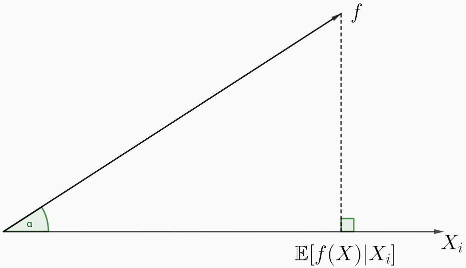
$$\operatorname{Var} f(\mathbf{X}) = \sum_{A \in \mathcal{P}(d)} \operatorname{Var} f_A(\mathbf{X}_A).$$

After renormalization:

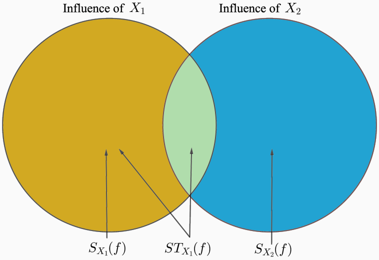
$$1 = \sum_{A \in P(d)} \overbrace{S_{\mathbf{X}_A}(f)}^{\text{Sobol' indices}} .$$

Sobol' indices, but simpler

In a nutshell: Sobol' indices $= \cos^2(\alpha)$.



What are Sobol' indices?



Two definitions (we denote by $\sim A := A^c$):

$$S_{X_i}(f) := \frac{\mathrm{Var}\,\mathbb{E}[f(\mathbf{X})|X_i]}{\mathrm{Var}\,f(\mathbf{X})}, \tag{1}$$

$$ST_{X_i}(f) := \sum_{s \ni X_i} S_{\mathbf{X}_s}(f) = 1 - S_{\mathbf{X}_{\sim i}(f)}. \tag{2}$$

Why Sobol' indices fall short.

Main **assumption** of the Hoeffding decomposition: **independent inputs** (not realistic).

Hence come the **extended Sobol' indices** [2] to differentiate:

- joint effects (*e.g.* $f(X_1, X_2) = X_1 \times X_2$) and
- intrinsic effect of an input variable with the others (*e.g.* $X_1 = g(X_2, \varepsilon)$ with ε some source of randomness).

Notation: $S_{X_i}(f)$ *is for independent inputs, otherwise we use* $Sob_{X_i}(f)$.

Some remarks on Sobol'-based indices

Sobol' indices		
	"Entanglement" between variables	Joined contributions
Sob_k	✓	✗
$SobT_k$	✓	✓
Sob_k^{ind}	✗	✗
$SobT_k^{ind}$	✗	✓

Table 1: Sobol' indices: what is taken into account and what is not.

We proved a **Central Limit Theorem** for Monte Carlo estimates of these quantities.

Welcome to the Fairness World

Group Fairness framework: we add a **sensitive feature** S (gender, ethnicity, etc...).

We want S **NOT** to be influent on the outcome $f(\mathbf{X}, S)$.

Note: Fairness through unawereness, i.e. "not looking at S" does not work.
Note bis: S multidimensional: notion of "intersectionality".

Classical Fairness definitions

Fairness definition	Binary formula
Statistical Parity	$\mathbb{P}(f(\mathbf{X}, S) = 1 \mid S = 0) = \mathbb{P}(f(\mathbf{X}, S) = 1 \mid S = 1)$.
Avoiding Disparate Treatment	$\mathbb{P}(f(\mathbf{X}, S) = 1 \mid \mathbf{X} = x, S = 0) = \mathbb{P}(f(\mathbf{X}, S) = 1 \mid \mathbf{X} = x, S = 1)$.
Equality of odds	$\mathbb{P}(f(\mathbf{X}, S) = 1 \mid Y = i, S = 0) = \mathbb{P}(f(\mathbf{X}, S) = 1 \mid Y = i, S = 1), i = 0, 1$.
Avoiding Disparate Mistreatment	$\mathbb{P}(f(\mathbf{X}, S) \neq Y \mid S = 1) = \mathbb{P}(f(\mathbf{X}, S) \neq Y \mid S = 0)$.

Table 2: Common fairness definitions and associated GSA measures

The link between GSA and Fairness

Theorem (B. & al., 2103.04613)
GSA measures define Fairness measures.

Fairness definition	GSA measure associated
Statistical Parity	$\mathrm{Var}(\mathbb{E}[f(\mathbf{X}, S)\|S]) \to Sob_S(f(\mathbf{X}, S))$
Avoiding Disparate Treatment	$\mathbb{E}[\mathrm{Var}(f(\mathbf{X}, S)\|X)] \to SobT_S(f(\mathbf{X}, S))$
Equality of odds	$\mathbb{E}[\mathrm{Var}(\mathbb{E}[f(\mathbf{X})\|S, Y]\|Y)] \to CVM^{ind}(f(\mathbf{X}, S), S\|Y)$
Avoiding Disparate Mistreatment	$\mathrm{Var}(\mathbb{E}[\ell(f(\mathbf{X}, S), Y)\|S]) \to Sob_S(\ell(f(\mathbf{X}, S), Y))$

Table 3: Common fairness definitions and associated GSA measures

Only good things can come from this...

Consequences of this theoretical link:

- **generalization of the fairness definitions to non-binary variables** (i.e. $S \in \{0,1\} \to S \in \mathbb{R}$),

Only good things can come from this...

Consequences of this theoretical link:

- **generalization of the fairness definitions to non-binary variables** (i.e. $S \in \{0,1\} \to S \in \mathbb{R}$),
- **fairness with respect to the predictor vs the error of the predictor** (i.e $GSA(f(\mathbf{X}, S))$ vs $GSA(\updownarrow(f(\mathbf{X}, S), Y))$),

Only good things can come from this...

Consequences of this theoretical link:

- **generalization of the fairness definitions to non-binary variables** (i.e. $S \in \{0, 1\} \to S \in \mathbb{R}$),
- **fairness with respect to the predictor vs the error of the predictor** (i.e $GSA(f(\mathbf{X}, S))$ vs $GSA(\updownarrow(f(\mathbf{X}, S), Y)))$,
- **definition of perfect and approximate fairness** (i.e.$GSA(f(\mathbf{X}, S)) \leqslant \varepsilon$, ε small).

Metamodels & Audits

What are metamodels?

- Sometimes, f is not accessible or is too costly.
- We can use an **approximation** $\widehat{f}$ of f.
- Question: if GSA_i is an index defined earlier, **how close is** $GSA_i(\widehat{f})$ **to** $GSA_i(f)$**?**
- Previous works: [1], [4]...

Result for Sobol'-based indices

We extend [4] to all the Sobol'-based indices defined earlier.

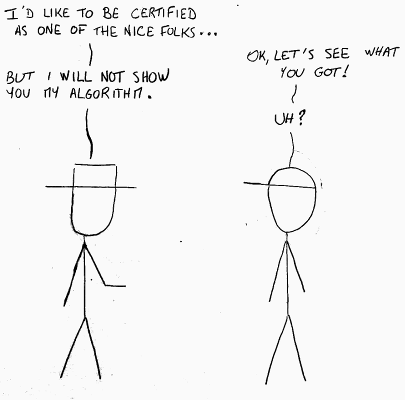
Table 4: Risk bounds for the various used GSA indices.

GSA index	Associated upper-bound
Extended Sobol' indices	$\frac{\mathbb{E}\left\|f-\widehat{f}\right\|_2^2}{\mathrm{Var}(f)}$
Extended Cramér-von-Mises indices	$\mathbb{E}\left\|f-\widehat{f}\right\|_2$
Shapley indices	$2 \times \frac{\mathbb{E}\left\|f-\widehat{f}\right\|_2^2}{\mathrm{Var}(f)}$

Next step: **asymptotic rates**, more if possible.

Metamodels & Audits

Translation in the Fairness world: audits!



- Corporations may be reticent about showing their algorithms for audits.
- Using GSA, we propose techniques for auditing using only metamodels.
- Warning: beware of "fair-washing"!

GSA 2: UQ strikes back!

Uncertainties, uncertainties everywhere...

What if **input distribution is not certain**? Most visual example:

$$\mathbb{P}_X = \varphi_\theta(x)dx, \theta \in \Theta.$$

Uncertainties, uncertainties everywhere...

What if **input distribution is not certain**? Most visual example:

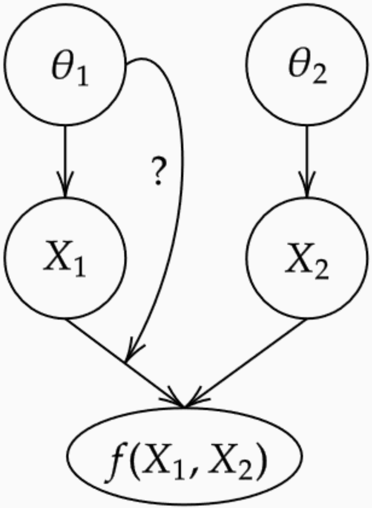
$$\mathbb{P}_X = \varphi_\theta(x)dx, \theta \in \Theta.$$

What happens to the GSA indices?

Second **level of uncertainty**: random distribution on θ.

$$GSA2_{X_i,\theta_i}(f) = GSA_{\theta_i}(GSA_{X_i}(f)).$$

Note: Initial idea from [3].



"Do you want a double loop or a single loop with this?"

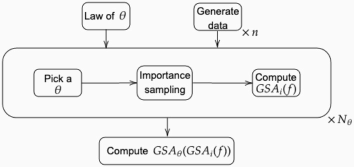


Figure 1: Workflow GSA2 in single loop

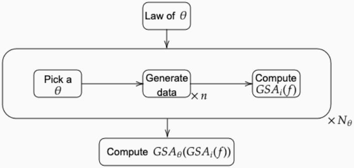
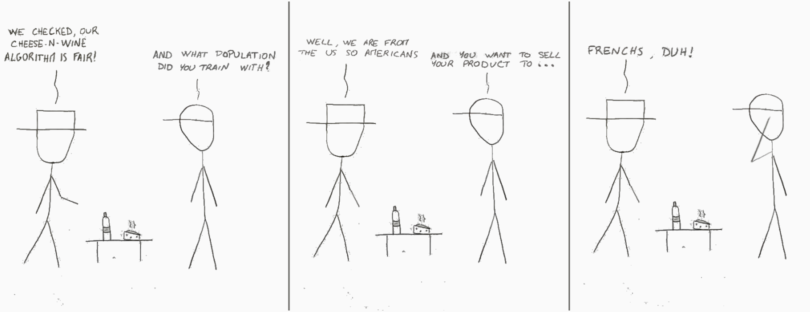


Figure 2: Workflow GSA2 in double loop

Pick'n'Freeze or **Chatterjee** estimators are consistent.

Fairness certification?

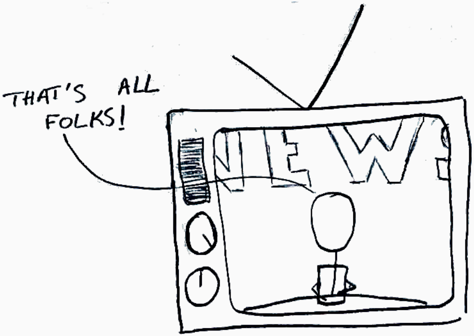


Training and real-life distributions can be different. We aim at **certifying fairness against distributional changes**.

To sum up

- **Link between GSA and Group Fairness**
- Behaviour of **Sobol'-based indices** under **metamodel** usage & **Fairness audits**.
- **Second-level GSA** and hints for **Fairness certification**.

Thanks for listening!



References

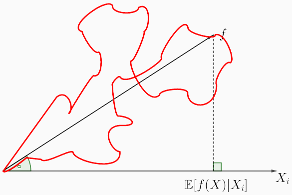
[1] Alexandre Janon, Maëlle Nodet, and Clémentine Prieur. "Uncertainties assessment in global sensitivity indices estimation from metamodels". In: *International Journal for Uncertainty Quantification* 4.1 (2014).

[2] Thierry A Mara, Stefano Tarantola, and Paola Annoni. "Non-parametric methods for global sensitivity analysis of model output with dependent inputs". In: *Environmental modelling & software* 72 (2015), pp. 173–183.

[3] Anouar Meynaoui, Amandine Marrel, and Béatrice Laurent. "New statistical methodology for second level global sensitivity analysis". In: *arXiv preprint arXiv:1902.07030* (2019).

[4] Ivan Panin. "Risk of estimators for Sobol'sensitivity indices based on metamodels". In: *Electronic Journal of Statistics* 15.1 (2021), pp. 235–281.

Annex: Cramér-von-Mises or "Sobol' indices on steroids"

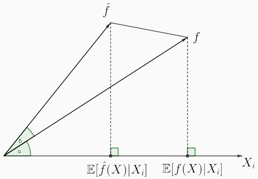


Some unusual definition of Cramér-von-Mises indices:

$$CvM_{X_i}(f) := \int Sob_{X_i}(\mathbb{1}_{f(.)\leqslant t}) \frac{\mathrm{Var}(\mathbb{1}_{f(.)\leqslant t})}{\int \mathrm{Var}(\mathbb{1}_{f(.)\leqslant t})dt} dt. \quad (3)$$

Note: Shapley indices are also related to Sobol' indices.

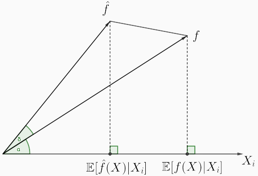
Annex:Proof in one picture



$$\cos^2(\alpha) - \cos^2(\alpha+\delta) = \frac{\cos(2\alpha)}{2} - \frac{\cos(2\alpha + 2\delta)}{2}$$

$$\sin(\theta)\sin(\varphi) = \frac{\cos(\theta-\varphi) - \cos(\theta+\varphi)}{2}$$

Annex:Proof in one picture



$$\cos^2(\alpha) - \cos^2(\alpha+\delta) = \frac{\cos(2\alpha)}{2} - \frac{\cos(2\alpha+2\delta)}{2}$$

$$\sin(\theta)\sin(\varphi) = \frac{\cos(\theta-\varphi) - \cos(\theta+\varphi)}{2}$$

$$|\cos^2(\alpha) - \cos^2(\alpha+\delta)| = \sin(2\alpha+\delta)\sin(\delta)$$