Shapley effects for target sensitivity analysis with correlated inputs: new insights

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

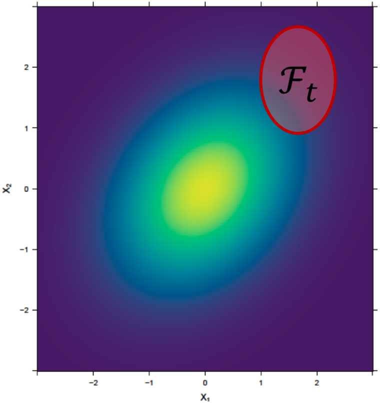
10th International Conference on SAMO
March, 16th 2022

Marouane Il Idrissi[1,2,3], Vincent Chabridon[1,3], Bertrand Iooss[1,2,3],

Target sensitivity analysis

Target sensitivity analysis (TSA) aims at **measuring the influence** of inputs on the **occurrence** of a failure event (Raguet and Marrel 2018).

- $X = (X_1, \ldots, X_d)$ is a random vector of **inputs**
- $Y = G(X)$ is the random **output** of a numerical model
- t is a **threshold**, such that $\{Y \geq t\}$ is a **failure event**
- $\mathcal{F}_t \in \mathbb{R}^d$ is the **failure domain**, i.e., $\mathcal{F}_t = \{y \in \mathbb{R}^d \mid G(y) \geq t\}$



The **variable of interest** is the **failure occurrence**:

$$\mathbb{1}_{\mathcal{F}_t}(x) = \mathbb{1}_{\{G(x) \geq t\}}(x)$$

with $p_t = \mathbb{P}(G(X) \geq t)$ the **failure probability**.

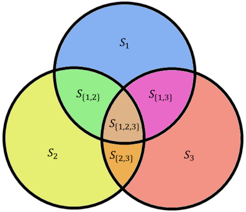
The **quantity of interest** is $\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right) = p_t(1 - p_t)$.

Target Sobol' indices

Whenever inputs are assumed to be **independent**, one can assess their influence through the **target Sobol' indices** (I., Chabridon, and Iooss 2021). For $A \subseteq \{1, \ldots, d\}$:

$$\text{T-}S_A = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \frac{\mathbb{V}(p_t^B)}{p_t(1-p_t)}$$

with $p_t^B = \mathbb{P}(G(X) > t \mid X_B)$, **conditional** probability failure given $X_B = (X_i)_{i \in B}$.



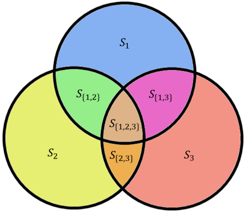
The **target Sobol' indices** are a tool to assess interactions in a reliability-oriented setting (Chabridon et al. 2021)

Target Sobol' indices

Whenever inputs are assumed to be **independent**, one can assess their influence through the **target Sobol' indices** (I., Chabridon, and Iooss 2021). For $A \subseteq \{1, \ldots, d\}$:

$$\text{T-}S_A = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \frac{\mathbb{V}(p_t^B)}{p_t(1-p_t)}$$

with $p_t^B = \mathbb{P}(G(X) > t \mid X_B)$, **conditional** probability failure given $X_B = (X_i)_{i \in B}$.



The **target Sobol' indices** are a tool to assess interactions in a reliability-oriented setting (Chabridon et al. 2021)

Model:

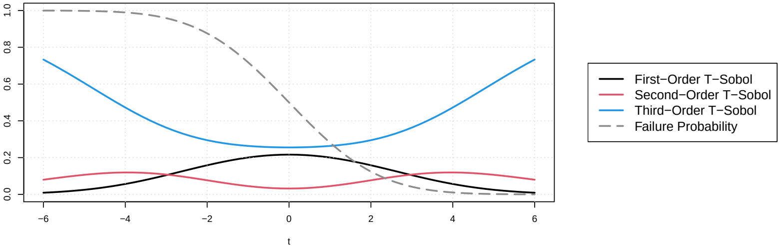
$$G(x) = X_1 + X_2 + X_3$$

Inputs:

$$X \sim \mathcal{N}(\mathbf{0}_3, I_3)$$

Variable of interest:

$$G(X) > t$$



Target Shapley effects

When inputs are **correlated**, (target) Sobol' indices can be **negative**: delicate interpretation (Da Veiga et al. 2021).

Solutions exists (Chastaing, Gamboa, and Prieur 2012; Mara and Tarantola 2012) but are often either **restrictive** or **challenging to estimate**.

Target Shapley effects

When inputs are **correlated**, (target) Sobol' indices can be **negative**: delicate interpretation (Da Veiga et al. 2021).

Solutions exists (Chastaing, Gamboa, and Prieur 2012; Mara and Tarantola 2012) but are often either **restrictive** or **challenging to estimate**.

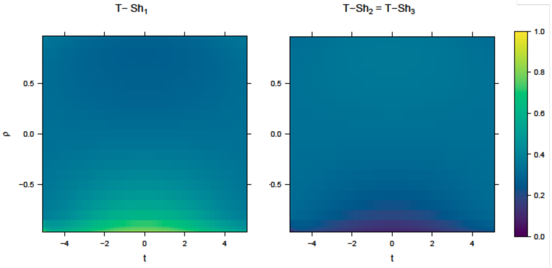
Owen (2014) proposed to use the **Shapley effects** (Shapley 1951). Transposition to TSA lead to the **target Shapley effects** $\mathrm{T}\text{-}Sh_i$.

They assess the influence of each input variable X_i, as an **aggregation of individual, interaction and dependence effects**, with the added properties:

- $\sum_{i=1}^{d} \mathrm{T}\text{-}Sh_i = 1$;
- $\mathrm{T}\text{-}Sh_i \geq 0$ for any $i \in \{1, \ldots, d\}$.

Subsequently, they can then be interpreted as **shares of variance**.

$$X \sim \mathcal{N}_3\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0 & 0\\0 & 1 & \rho\\0 & \rho & 1\end{pmatrix}\right)$$



Cooperative game theory

In a nutshell, cooperative game theory can be summarized as **"the art of cutting a cake"**.

Given a **set of players** $D = \{1, \dots, d\}$, who produces a **quantity** $v(D)$:

How can one allocate shares of $v(D)$ among the d players?

The **"cake cutting process"** is often described through **axioms** (i.e., desired properties), and results in an **allocation**.

Formally, a cooperative game is denoted by (D, v) where D is the **set of all players**, and $v : \mathcal{P}(D) \to \mathbb{R}$ a **value function**, mapping every possible subset of players to a real value.

Shapley values

Harsanyi dividends allow for an intuitive equivalent characterization. Given a cooperative game (D, v), its Harsanyi dividends are defined, for any subset of players as:

$$\mathcal{D}_v(\{i\}) = v(\{i\}), \quad \forall i \in D$$
$$\mathcal{D}_v(A) = v(A) - \sum_{B \subset A} \mathcal{D}_v(B), \quad \text{recursively}, \quad \forall A \subseteq D$$

or more generally, for any $A \in \mathcal{P}(D)$:

$$\mathcal{D}_v(A) := \sum_{B \subseteq A} (-1)^{|A \setminus B|} v(B).$$

Shapley values can then be defined, for every player $i \in D$, as:

$$Shap_i = \sum_{A \in \mathcal{P}(D) : i \in A} \frac{\mathcal{D}_v(A)}{|A|}$$

The Shapley values **redistributes equally** the dividends due to a coalition **among the players that composes it**. This characterization is **equivalent** to the other representations (Harsanyi 1982).

Cooperative games and TSA

An analogy can be made between **players** and **inputs**. The chosen value function, for the target Shapley effects is:

$$v(A) = \frac{\mathbb{V}(p_t^A)}{p_t(1-p_t)} =: \text{T-}S_A^{clos}$$

which are the **closed target Sobol' indices**, and can be computed **without an independence assumption**. The cooperative game formed by $(D, \text{T-}S^{clos})$ allows for the following Harsanyi dividends:

$$\mathcal{D}_{\text{T-}S^{clos}}(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \frac{\mathbb{V}(p_t^B)}{p_t(1-p_t)} = \text{T-}S_A$$

and in turn, the **target Shapley effects** are the Shapley values of $(D, \text{T-}S^{clos})$ and can be written, for any $i \in \{1, \dots, D\}$, as (Plischke, Rabitti, and Borgonovo 2021):

$$\text{T-}Sh_i = \sum_{A \in \mathcal{P}(D) : i \in A} \frac{\text{T-}S_A}{|A|}$$

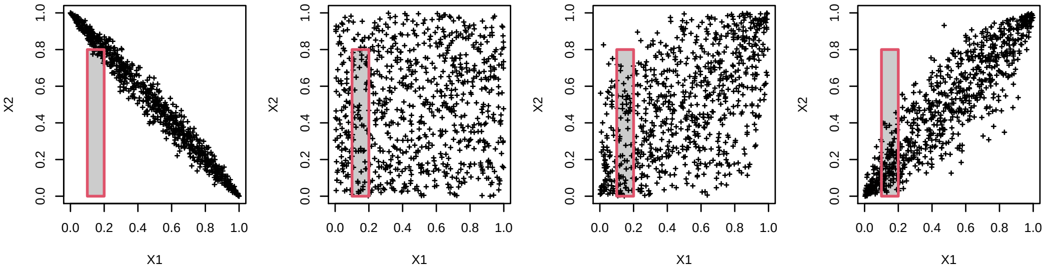
This formulation allows to assess **effects of a particular variable subset order**.

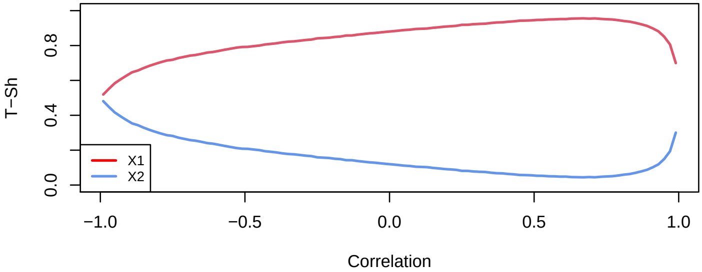
Estimating the target Shapley effects

In order to compute the target Shapley effects, one needs to estimate the **closed target Sobol' indices**, for every possible subset of variables $A \in \mathcal{P}(D)$. Different estimation strategies can be considered.

- Sampling strategies (requires the ability to sample from the **conditional laws** of the inputs)
  - Crude Monte Carlo (I., Chabridon, and Iooss 2021)
  - Monte Carlo/Pick-Freeze with Importance Sampling (Demange-Chryst, Bachoc, and Morio 2022)
- Given-data strategies (only when an i.i.d. sample is available)
  - Nearest-Neighbor procedure (Broto, Bachoc, and Depecker 2020)

Failure rectangle and correlated uniform inputs





Simplified flood model

Simplified flood use-case (Iooss 2011): river **water level** model, and occurrence of an **industrial site flood** when $G(X) > t$.

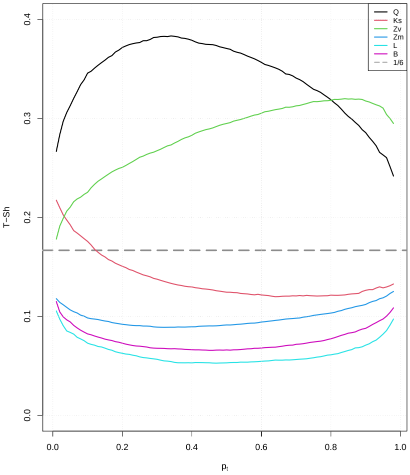
$$G(X) = Z_v + \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{\frac{3}{5}}$$

Input	Description	Distribution
Q	max flow rate	Gumbel$(1013, 558)$ trunc. $[500, 3000]$
K_s	Strickler coefficient	Normal$(30, 7)$ trunc. $[15, +\infty)$
Z_v	downstream level	Triangular$(49, 50, 51)$
Z_m	upstream level	Triangular$(54, 55, 56)$
L	length	Triangular$(4990, 5000, 5010)$
B	width	Triangular$(295, 300, 305)$

Correlation structure from Chastaing, Gamboa, and Prieur (2012):

$$Cov(Q, K_s) = 0.5, Cov(Z_v, Z_m) = 0.3, Cov(L, B) = 0.3$$

Nearest-neighbor estimation with `n.knn=3`, on a 2×10^5 i.i.d. sample.



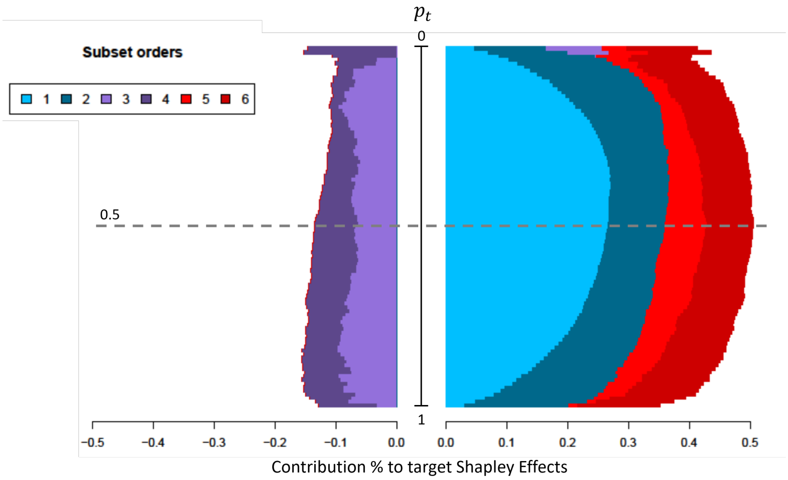
T-Sh order decomposition: maximum annual flow rate

Target Shapley effects:

$$\text{T-}Sh_i = \sum_{A \in \mathcal{P}(D) : i \in A} \frac{\text{T-}S_A}{|A|}$$

Contribution percentage to target Shapley effects, **for subsets of order** $k \in D$:

$$\frac{1}{\text{T-}Sh_i} \sum_{A : |A| = k, i \in A} \frac{\text{T-}S_A}{k}$$



Conclusion and perspectives

Harsanyi dividends play a central role in **understanding of the Shapley values**.

In GSA, they allow to assess **shares of Shapley effects** due to **subsets of variables**.

In TSA, they allow to better understand the empirical **behavior** of the **target Shapley effects** in restrictive settings.

Shapley values are a **particular choice of allocation**. Other choices may lead to better suited effects, depending on the goal and the context of the sensitivity analysis (Hérin et al. 2022).

References i

Broto, B, F Bachoc, and M Depecker. 2020. "Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution" [in en]. *SIAM/ASA Journal on Uncertainty Quantification* 8, no. 2 (January): 693–716. ISSN: 2166-2525, accessed December 2, 2020. https://doi.org/10.1137/18M1234631. https://epubs.siam.org/doi/10.1137/18M1234631.

Chabridon, V., M. Balesdent, G. Perrin, J. Morio, JM. Bourinet, and N. Gayton. 2021. "Global Reliability-oriented Sensitivity Analysis under Distribution Parameter Uncertainty" [in en]. In *Mechanical Engineering under Uncertainties,* 237–277. John Wiley & Sons, Ltd. ISBN: 978-1-119-81763-5, accessed March 15, 2022. https://doi.org/10.1002/9781119817635.ch7. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119817635.ch7.

Chastaing, G., F. Gamboa, and C. Prieur. 2012. "Generalized Hoeffding-Sobol Decomposition for Dependent Variables - Application to Sensitivity Analysis." *Electronic Journal of Statistics* 6:2420–2448.

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.

Demange-Chryst, J., F. Bachoc, and J. Morio. 2022. "Shapley effect estimation in reliability-oriented sensitivity analysis with correlated inputs by importance sampling." ArXiv: 2202.12679, *arXiv:2202.12679* [*math, stat*] (February). http://arxiv.org/abs/2202.12679.

Harsanyi, John C. 1982. "A Simplified Bargaining Model for the n-Person Cooperative Game" [in en]. In *Papers in Game Theory,* edited by John C. Harsanyi, 44–70. Theory and Decision Library. Dordrecht: Springer Netherlands. ISBN: 978-94-017-2527-9. https://doi.org/10.1007/978-94-017-2527-9_3. https://doi.org/10.1007/978-94-017-2527-9_3.

Hérin, M., M. Il Idrissi, V. Chabridon, and B. Iooss. 2022. "Proportional marginal effects for sensitivity analysis with correlated inputs." In *Proceedings of the 10th International Conference on Sensitivity Analysis of Model Output* (*SAMO 2022*). Tallahassee, Florida, USA.

I., M., V. Chabridon, and B. Iooss. 2021. "Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs" [in en]. *Environmental Modelling & Software* 143 (June): 105115. ISSN: 1364-8152. https://doi.org/10.1016/j.envsoft.2021.105115.

Iooss, B. 2011. "Revue sur l'analyse de sensibilité globale de modèles numériques" [in fr]. Number: 1, *Journal de la Société Française de Statistique* 152, no. 1 (January): 3–25. ISSN: 2102-6238. http://journal-sfds.fr/article/view/53.

Mara, T., and S. Tarantola. 2012. "Variance-based sensitivity indices for models with dependent inputs." *Reliability Engineering & System Safety* 107:115–121.

Owen, A. B. 2014. "Sobol' Indices and Shapley Value" [in English]. *SIAM/ASA Journal on Uncertainty Quantification* 2, no. 1 (January): 245–251. ISSN: 2166-2525, accessed December 2, 2020. https://doi.org/10.1137/130936233.

Plischke, E., G. Rabitti, and E. Borgonovo. 2021. "Computing Shapley Effects for Sensitivity Analysis." Publisher: Society for Industrial and Applied Mathematics, *SIAM/ASA Journal on Uncertainty Quantification* 9, no. 4 (January): 1411–1437. https://doi.org/10.1137/19M1304738. https://epubs.siam.org/doi/abs/10.1137/19M1304738.

Raguet, H., and A. Marrel. 2018. *Target and Conditional Sensitivity Analysis with Emphasis on Dependence Measures.* Research Report. CEA, DEN, DER, January. https://hal.archives-ouvertes.fr/hal-01694129.

References iii

Shapley, L. S. 1951. *Notes on the n-Person Game – II: The Value of an n-Person Game* [in English]. Research Memorandum ATI 210720. Santa Monica, California: RAND Corporation, August. https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.pdf.

Thank you for your attention!

Any question?

Möbius inverse

Let $\mathcal{P}(D)$ be the set of subsets of a finite set D (i.e., its powerset). One has that the Möbius function becomes:

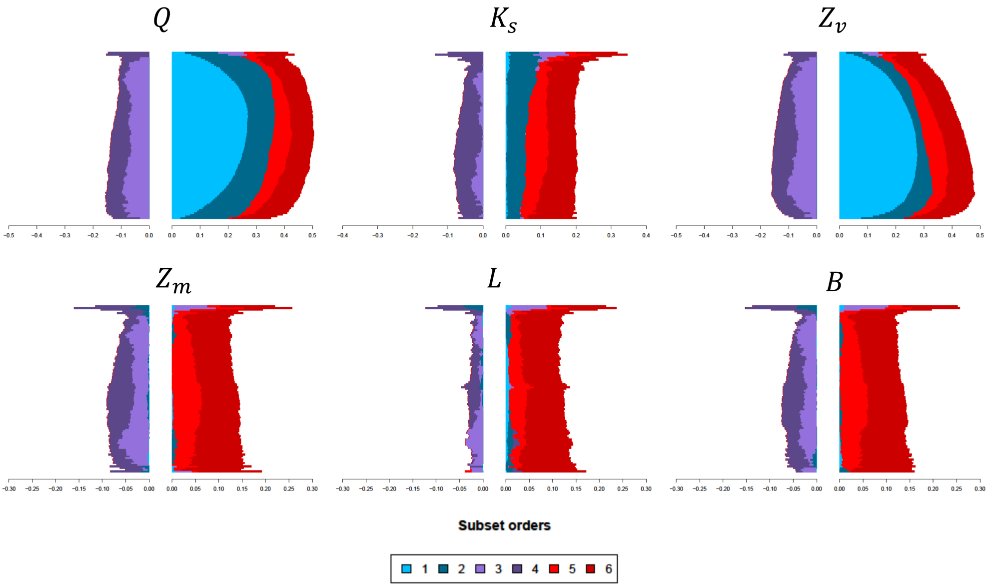
$$\mu(B, A) = (-1)^{|A \setminus B|}$$

for any pair of subsets $A, B \subseteq D$, such that $B \subseteq A$. The Möbius inversion formula then states that for any two functions f, g defined on $\mathcal{P}(D)$:

$$f(A) = \sum_{B \subseteq A} g(B) \quad \text{if and only if} \quad g(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} f(B)$$

moreover, this is called the **inclusion-exclusion principle**.

Flood model orders decomposition



Estimation scheme : Monte Carlo

The estimation of the target Shapley effects can be split into two steps:

- **Step #1**: estimation of the *conditional elements*, i.e., the estimation of $\mathrm{T\text{-}S}_A$ or $\mathrm{T\text{-}E}_A$ for all $A \in \mathcal{P}_d$;
- **Step #2**: an *aggregation procedure*, i.e., a step to compute the $\mathrm{T\text{-}Sh}_j$ by plugging in the previous estimations of Step #1.

In order to estimate a conditional element $\mathrm{T\text{-}S}_A$, one needs to draw several i.i.d. samples:

- an i.i.d. sample of size N drawn from P_X and denoted by $(X^{(1)}, \ldots, X^{(N)})$;
- another i.i.d. sample of size N_v drawn from P_{X_A} and denoted by $(X_A^{(1)}, \ldots, X_A^{(N_v)})$;
- for each element $X_A^{(i)}$, $i = 1, \ldots, N_v$, a corresponding sample of size N_p drawn from $P_{X_{\overline{A}}|X_A}$ given that $X_A = X_A^{(i)}$ and denoted by $(\widetilde{X}_i^{(1)}, \ldots, \widetilde{X}_i^{(N_p)})$.

Then, the Monte Carlo estimator of $\mathrm{T\text{-}S}_A$ can be defined as:

$$\widehat{\mathrm{T\text{-}S}}_{A,\mathrm{MC}} = \frac{\sum_{i=1}^{N_v} \left(\frac{1}{N_p} \sum_{j=1}^{N_p} \mathbb{1}_{\mathcal{F}_t}(\widetilde{X}_i^{(j)}, X_A^{(i)}) - \widehat{p}_t^Y \right)^2}{(N_v - 1)\widehat{p}_t^Y (1 - \widehat{p}_t^Y)} \tag{1}$$

Estimation scheme : Nearest-neighbor

Let $\left(X^{(1)}, \ldots, X^{(N)}\right)$ be an i.i.d. sample of the inputs X and $A \in \mathcal{P}_d \setminus \{\emptyset, [1:d]\}$. Let $k_N^A(l, n)$ be the index such that $X_A^{\left(k_N^A(l,n)\right)}$ is the n-th closest element to $X_A^{(l)}$ in $\left(X_A^{(1)}, \ldots, X_A^{(N)}\right)$. Note that, if two observations are at an equal distance from $X_A^{(l)}$, then one of the two is uniformly randomly selected. Finally, one can define an estimator of the equivalent value function:

$$\widehat{\text{T-E}}_{A,\text{KNN}} = \frac{\sum_{l=1}^{N} \left(\frac{1}{N_s - 1} \sum_{i=1}^{N_s} \left[\mathbb{1}_{\mathcal{F}_t} \left(X^{\left(k_N^{\bar{A}}(l,i)\right)} \right) - \frac{1}{N_s} \sum_{h=1}^{N_s} \mathbb{1}_{\mathcal{F}_t} \left(X^{\left(k_N^{\bar{A}}(l,h)\right)} \right) \right]^2 \right)}{N \widehat{p}_t^Y (1 - \widehat{p}_t^Y)}. \tag{2}$$

Under some mild assumptions, Broto, Bachoc, and Depecker 2020 showed that this estimator does asymptotically converge towards T-E_A.